University of Tartu
Faculty of Science and Technology
Institute of Ecology and Earth Sciences
Department of Geography

Master's thesis in Geoinformatics for Urbanised Society (30 ECTS)

Exploring Social Networks and Spatial patterns of Information Dissemination in Passive Mobile Positioning Data

Lika Zhvania

Supervisors:
Ph.D. Anto Aasa
Ph.D. Anniki Puura

Tartu 2023

Abstrakt

Sotsiaalvõrgustikud ja informatsiooni levitamise ruumilised mustrid passiivsete mobiilpositsioneerimise andmete põhjal

Uurimistöös on andmepõhise lähenemise kaudu analüüsitud telefonikasutajate vahelist sotsiaalvõrgustikku Eestis mobiilpositsioneerimise andmete alusel. Uurimistöö täiendab olemasolevaid sotsiaalvõrgustike alaseid teadmisi analüüsides ja kirjeldades sotsiaalvõrgustike ruumilisi mustreid. Uurimistöö kasutab võrgustiku- ja ruumianalüüsi tööriistu, et uurida telefonikasutajate suhtlusvõrgustike mustreid ja struktuuri, mis omakorda aitab suunata potentsiaalsete uurimisteemade püstitamist tulevikus. Uurimistöö keskseteks teoreetilisteks lähtekohtadeks on võrgustike kesksusnäitajad ja inimestevaheliste sotsiaalsete sidemete tugevus, mis põhineb kõnepaaride suhtlusaktiivsusel. Võrgustike kesksusnäitajate kaudu on uurimistöös tuvastatud võrgustiku kesksete liikmete ruumilist paiknemist asustusüksuse täpsusega. Võrgustiku liikmete kõrget kesksust võib seostada suure mõjuvõimuga kogu võrgustikus. Suhtlusvõrgustike sidemete tugevuse uurimine abil leiti nõrgad ja tugevad ühendused võrgustiku liikmete vahel ja seejärel tuvastati need nõrgad ühendused, millel on oluline roll erinevate suhtlusgruppide omavahelisel ühendamisel. Lisaks suhtlusvõrgustike teadmiste, käitumise ja ressursside levitamise seisukohast oluliste asustusüksuste tuvastamisele leiti ka kõrgema haavatavusega ühendused, mis võib viidata piiratumale ligipääsule uuele informatsioonile.

Märksõnad: sotsiaalne võrgustik, mobiilpositsioneerimise andmed, võrgustikuanalüüs, ruumianalüüs, Eesti

CERCS kood: S20 – Sotsiaalne geograafia

Abstract

Exploring Social Networks and Spatial Patterns of Information Dissemination in Passive Mobile Positioning Data

This study represents the data-driven approach to analysing the social network of call pairs in Estonia by utilising the mobile phone dataset. The study aims to bring new knowledge to the field by deriving the spatial context from the mobile phone dataset and addressing it to social network analysis. On the one hand, the nature of the study is exploratory, implementing data utilisation and graph and spatial analysis tools to explore the structure and patterns of the social network of call partners, aiding the understanding of the potential prospects of further research. On the other hand, the focus is on the central research questions, which lie in two parts: network centrality and the strength of the dyadic ties. By applying the network centrality concept, the network's essential actors associated with power and leadership were identified based on the settlement types and the spatial context. Analysing the strength of the ties resulted in detecting the weak and strong links and identifying the weak ties with importance in bridging different social networks or groups based on settlement types and spatial distribution. Besides identifying the important channels for spreading diverse information, knowledge, behaviour or resources, the vulnerable ties are also detected, indicating to the actors with limited sources of information.

Keywords: social network, mobile phone data, network analysis, spatial analysis, Estonia

CERCS code: S230 – Social geography

Table of Contents

Table of Figures

1. Introduction

Social interactions are the source for exchanging ideas, knowledge, and information and spreading influence. The patterns of information dissemination or access to the novelty depends on the forms of social networks and their channels. Understanding social network structure might help identify isolated and vulnerable groups, detect mobility patterns or address the patterns in various social studies.

Many researchers have contributed to the study of social networks. The facets, such as the strength of the ties, are investigated, as well as identifying the patterns based on attributes such as age and sex. The importance of weak ties is also addressed for gaining social benefits. Social network analysis is implemented in social studies, such as mobility, tourism and segregation. The focus on studying personal social networks concerning mobility has also attracted the researchers' attention.

Even though attention to social network studies is significant, there are gaps that this thesis aims to fill and bring new knowledge to the field by focusing on the spatial aspect of the study from the perspective of the relationship between settlements based on the settlement types.

The study aims to analyse the social network of call pairs concerning the settlement types of the network actors and identify the spatial patterns based on Estonia. The research consists of two main parts: a data-driven approach to utilise the data and implement the tools to conduct graph and spatial analysis and a focus on the main research questions to identify the structure of the network.

Research questions are set as follows:

1. What are the spatial patterns and characteristics of the network centrality? What are the patterns of the important actors by the settlement types?

2. What are the spatial patterns of the strength of social ties? Identify the strong and weak ties, and detect the essential links.

The study utilises mobile phone (Call Detail Records, CDRs) data. The context of the research is Estonia. The data covers the month of February 2018.

The theory of the study consists of four units. The first subchapter focuses on the broader field of network science and its importance in studying social networks. While the second subchapter briefly discusses the field of the social network. The third subchapter describes the social network concepts corresponding to the main research questions. The final subchapter is an overview of the recent studies on social networks.

2. Theory

2.1. Network science and its importance for social network analysis

Social networks have existed since humankind when societies were still fragmented and earthbound, long distances dispersed communities, and news and ideas travelled and spread on foot (Barabási, 2003). Since then, till now, the structure of social networks has changed significantly (Rainie & Wellman, 2014), though it has become trendy to study social networks only recently due to recent technological development. The twenty-first century brought a new era of technology, allowing the creation of instruments and tools for data assembly, sharing, analysis and implementation in studies, making human behaviour observed and measurable phenomena. The technological improvement allows gathering non-traditional data like people's everyday actions and communications and opening prospects for scientific research. Thus, existing and potential digital data sources are leading to transforming social sciences (Lazer et al., 2021). However, it is challenging to understand social networks due to their complexity. Social networks are complex systems of individuals, kinships, friendships and other relationships and interactions. Like in other networks, the characteristics of the complexity in social networks lie in the interaction of compound parts. Understanding the nature of complex systems in the connected era requires a scientific description that demands the need for network science (Watts, 2004).

Network science as a separate discipline has emerged in the twenty-first century (Barabási & Pósfai, 2016) and the field is forming and shaping (Brandes et al., 2013). Although network science has emerged as a separate field (Barabási & Pósfai, 2016), it embeds in traditional disciplines (Brandes et al., 2013). It refers to studying the networks of complex systems, including exploring human social phenomena. Due to the characteristics of networks, like interconnectivity, networks bear benefits and vulnerabilities. One of the essential characteristics of network science is its empirical and data-driven nature, focusing on data, function and utility (Barabási & Pósfai, 2016).

For social network analysis, it is crucial to understand its broader field, network science, which offers data collection methods, innovative mathematical techniques and predictive theories (Hansen et al., 2020). The structure of networks, in general, consists of nodes and links, also referred to as vertices and edges in graph theory. In social network analysis, nodes are interpreted as actors (humans or organisations) and lines as relations between actors. They can be drawn as social networks or graphs (Bruggeman, 2008). By understanding key properties of network science, it is conceivable to measure core parameters such as the degree and weight of given social networks and describe the main patterns. At the same time, more comprehensive algorithms allow for identifying clusters in social networks, studying connectedness between individuals or groups in social networks, or detecting communities (Barabási & Pósfai, 2016). Network analysis encourages studying the representation of network structure, individual elements, and pair-wise relationships between elements (Brandes et al., 2013). Analysing dyadic variables is especially important in social network research.

2.2. The field of social network

The social network field is relatively new, though the social network community has emerging roots in the 1940s (Barabási & Pósfai, 2016). Moreover, the problems that are still active have been addressed earlier. In 1929, in his book, "Everything Is Different", Hungarian poet and writer Frigyes Karinthy included the story called "Chains" ("Láncszemek"), which is the basis of the well-known concept of six degrees of separation (Barabási, 2003). Karinthy emphasises the crucial change in Earth regarding becoming tiny and shrunk due to the rising rhythm of physical and verbal communication (Karinthy, 1929). The shrinking in size prompted the enlargement of social networks and channels. Karinthy claims in his story that two individuals in the chain are linked with each other at most by five links (Karinthy, 1929). Later, the concept developed as six degrees of separation, indicating the theory of two random individuals being linked by no more than six intermediaries in the friendship chain (Zhang & Tu, 2009).

A noteworthy example in the social network is the approach to understanding the influence of links in social networks. J. L. Moreno created a method called "Sociometry" to study the interpersonal connections of individuals and groups (Hale, 2009), which he implemented in a real-case problem known as epidemic runaways of girls at the Hudson School in 1932 (Borgatti et al., 2009). Moreno, with Helen Jennings, tried to map and graphically represent social networks. Moreno explained the event as the impact of social influence and ideas flown to channels through links (Hale, 2009).

Social network studies have progressed further since the 1940s and 1950s by putting the conceptual focus on groups and social circles in the network based on graph theory, accompanied by designing a program to study the network structures (Borgatti et al., 2009). The graph theory had its roots in 1735 (Barabási & Pósfai, 2016) when mathematician Euler offered the solution to the famous problem of "Bridges of Königsberg" by using graphs (Wilson, 1996). Later, graph theory problems were tackled with randomness in graphs by mathematicians Erdős and Rényi (Barabási, 2003), nowadays known as Erdős-Rényi random graph model used in network analysis. Even though graph theory was established as a mathematical tool, it has been used widely in various fields, including social sciences (Wilson, 1996). The graph theory concept is also vastly used in social media (Chakraborty et al., 2018), one of the most significant digital data sources for social network analysis.

In the 1950s, views started circulating on social stratification, the subject of discussion among sociologists, and influence, the subject of discussion among political scientists, by combining these two concepts regarding studying social networks. The assertion that influence is an essential aspect of reaching channels in social networks and the higher number of channels ensuring better connectedness aroused the studies of human contact nets (Pool & Kochen, 1978). Studies of empirical estimates of acquaintanceship parameters by Pool and Kochen, based on the case of the population of the United States, suggest that, at most, seven intermediaries are required to link two random individuals (Pool & Kochen, 1978), and at least 50% of pairs could be linked by the chain of no more than two intermediaries (Borgatti et al., 2009). These estimates tackled the problem

known as the small world phenomenon or the "six degrees of separation", addressing the issue of short paths in social networks and prompted the first significant empirical study by Milgram (Easley & Kleinberg, 2010).

Milgram (1967) developed two philosophical views around the small-world problem. First, two random individuals might be linked through intermediate acquaintances due to the intersection of the social groups via connecting links. Second, two random individuals might not be linked due to circles of acquaintances prone to not intersect with each other. Based on these views, characteristics in groups of acquaintances, such as concentration inside the circle and forming orbits, cause the isolation of groups, while connecting links are valuable for linking different group members with each other. Milgram (1967) argued that the number of acquaintances defines the number of lines between two random individuals. He conducted an experiment to estimate the number of intermediary links between two random people. The experiment contained a complex network of two hundred million points (people) and tested the shortest paths connecting any given two points from this network. It was assumed that a person's average number of acquaintances is roughly 500. The small-world experiment demonstrated that the number of intermediate acquaintances in a chain varies from two to ten, and the median is five (Milgram, 1967). A similar experiment about the small-world problem was carried out again by Stanley Milgram and Jeffrey Travers in 1969. This time, the focus was on variation in the procedure, particularly involving three distinctive subpopulations as the "starting population" of the experiment. Travers & Milgram (1969) attempted to assess the impacts of geographical distance and the importance of the target's occupational group. According to the experiment's outcomes, the estimated mean for the number of intermediaries required to link starters and targets is 5.2 (Travers & Milgram, 1969). From the experiment, the following was learned: social proximity depends on geographical proximity since the experiment revealed the importance and impact of geographical distance on the number of intermediaries in the chain; attributes such are age, sex, and occupation impacted the contact selection among participants of the experiment (Travers & Milgram, 1969).

Barabási (2003) argues that the number of separators between two random individuals has decreased from what Karinthy and Milgram estimated, and nowadays, it is close to three intermediaries. Karinthy's (1929) indication of the shrinking of the Earth is even more actual nowadays, considering the increased rhythm of social life due to technological development.

It is essential to briefly discuss the societal changes and causing factors to understand social network structure and its complexity.

The structure of societies has changed significantly due to various factors. Heretofore, societies were fragmented and earthbound, and the distance between them was ample (Barabási, 2003), shaping the social structure in small groups and communities. Over time, social networks were encouraged to modify for various reasons, while transportation development supported the most critical aspect, which affected the far-flung development of social ties (Rainie & Wellman, 2014). Nowadays, the structure of

social networks has turned from groups into individualised networks where "networked individuals have partial membership in multiple networks and rely less on permanent memberships in settled groups". According to Rainie & Wellman (2014), three revolutions, Social Network Revolution, Internet Revolution and Mobile Revolution, are the grounds for shifting individuals from being embedded group members into being connected individuals.

Moreover, the rapid growth of computer-supported information and communication technologies (ICTs) not only rose personal connectivity but also changed the ways of communication and interaction, allowing people to be part of various social networks. Even though complex social networks existed before, technological developments in communication played an essential role in structuring social networks, notably revealing the affordability of social relations over long distances due to the "social affordance" of technology (Wellman, 2001). Furthermore, communication tools create new patterns of connectedness since individuals separated due to physical distance are brought socially close (Licoppe & Smoreda, 2005).

2.3. Social network concepts and theories

Dyadic relationships create networks, and networks, on the other hand, form the social structure through the connected actors. There are two primary types of social circles. First, characterised by emotional closeness and short paths of intermediaries between individuals and is represented chiefly in family forms, and second, represented in rational circles such as partnerships, business relationships, when the circle is based on similar interests and "homophily" (Gamper, 2022). Homophily refers to the homogeneity of the group when the actors with similar attributes tend to communicate with each other more frequently than with dissimilar actors, thus making the social space of a group localised (McPherson et al., 2001). In the social network of call pairs, localisation might be discussed regarding the information spread through the homogeneous groups.

Two main concepts, network centrality and the strength of ties, are discussed to respond to the central research questions of the study.

2.3.1. Network centrality

Network centrality is one of the primary attributes of the structure of social networks and refers to identifying the important actors in the network. According to Freeman (Freeman, 1978a), studies on centrality have revealed its role in groups regarding group efficiency in problem-solving. The concept of centrality is widely applied in various applications, including matters such as political integration, urban development, the structure of organisations, inter-organisational relations, etc.

Centrality measures identify essential actors in the network, though, in general, the importance of actors should not necessarily be related to power. Bonacich discusses the relationship between centrality and power and claims that the assumption that centrality produces power has weakened in the case of particular types of networks, such as

exchange networks, for instance. Moreover, in particular cases such as trading partners, connectedness to actors with low centrality degrees might ensure power to the connected actor(s) (Bonacich, 1987). Though in social networks, oppositely, centrality is viewed as power, and high centrality position might be linked with social roles such as leadership (Leavitt, 1951; Berkowitz, 1956; Shaw, 1964; cit. Bonacich, 1987).

Centrality measures as the mechanism to identify leaders based on their position in a network are discussed regarding the two-step flow of communication hypothesis (Liu et al., 2017). The hypothesis refers to two levels of influence: mass media to opinion leaders and leaders to their social circles (Lazarsfeld et al., 1944; cit. Liu et al., 2017). The impact of personal influence on social networks is observed in such important decisions as voting. The significant point is that the position in the network, possibly together with special attributes, constructs particular individuals as leaders that finally ensure the dissemination of the information and the influence (Liu et al., 2017).

The centrality measures are responsible for structural measurement and analysis. (Carrington et al., 2005). Freeman (1978) discusses the three distinct structural properties of centrality for an actor in a network: degree, the point's (actor) position in a graph with the utmost possible degree; betweenness, the point with position between the most significant number of other points; closeness, when the point is located at the minimum distance from all other points (Freeman, 1978a). However, additionally, various measurements exist for centrality analysis. The consideration of different measurements is the functionality they are based on and the trajectories of the paths among nodes that might impact the interpretation of the measured entities or poor outcomes (Borgatti, 2005). The path's trajectory depends on the type of traffic flow, which might be the flow of transportation, finances, gossip, infection or other. The characteristics of the traffic define the network structure (Borgatti, 2005).

The degree is the most straightforward measure of centrality, counting the number of edges for each node and referring to the number of ties of the actor (Hoffman, 2021). The degree measurement depends on the mode, defined as in-degree, out-degree and all-degree, referring to the direction of the link. In this research, in-degree corresponds to incoming and out-degree to outgoing calling activities between a calling pair.

2.3.2. Strong and weak ties

The essential aspect of social ties is the strength, how strongly or weakly the actors are connected. From the perspective of dissemination of information, ideas, influence, support etc., it is especially crucial to explore the strength of the dyadic connections in the social network.

The characteristic of the strength of ties is a combination of factors such are the frequency of time, emotional attachment, intimacy and reciprocity (Granovetter, 1973). It is challenging to assume strength of a tie is strong, weak or absent. However, acquaintanceship and less contribution to the mentioned aspects might elucidate social ties as weak. In contrast, kinship and close friendship ties contribute more to the

relationship and shape it as a strong dyad. Nevertheless, weak ties have the strength that lies in disseminating information and influence. While in a group of dyads with more or less equally strong ties, the same knowledge and information are circulated, weak ties linked with other groups channel the novelty in a group (Granovetter, 1973).

Weak ties might be discussed from the concept of bridging in network science. The bridge is a single link that, if appropriately placed, connects two components in the network, and when the link is cut, the network becomes disjointed (Barabási & Pósfai, 2016). Bridgers ensure the spreading of information inside and outside a cluster of social groups. Thus, as a source of different information, the bridges are essential; however, the ties that stay inside the group and are called bonding ties are also vital for building internal "trust, efficiency, and solidarity" (Rainie & Wellman, 2014). Absent of bridges between different groups are prone to impediments in understanding or communicating with each other. Thus the roles of bridging ties and bonding ties, whose structural roles are "cosmopolitans" and "locals" (Merton, K., 1957; cit. Rainie & Wellman, 2014), are essential for societies. As a result of bridging "locals" and "cosmopolitans, processes that are simple by nature and occur at the level of individual actors and ties transform towards complexity that finally impacts society (Easley & Kleinberg, 2010).

Technological development, especially the internet, helps preserve weak ties over distance (Rainie & Wellman, 2014), supporting the perspective of analysing the strength of the ties with the spatial context.

2.4. Recent studies of social networks

As stated earlier, technological communication development proposed new data sources for researchers, which is widely grasped in many fields of social sciences. That also tackled social network analysis regarding the various facets of social studies. This chapter is devoted to a brief overview of the studies related to social networks, primarily based on mobile data utility. The purpose of this chapter is to address the aspects of studies regarding social networks, demonstrate the importance of using mobile data, and reveal the gaps in the field.

One of the key aspects of social network studies is the strength of social interactions. The frequency of social interaction between dyads is essential for the prevention of the decay of the relationship since social relationships are dynamic, and actively keeping ties requires social investment, which applies to not only strong but also weak ties (Dindia & Canary, 1993; Burt, 2000; cit. Roberts & Dunbar, 2011). The study of the effects of kinship, network size, and emotional closeness based on the social networks of 251 women revealed the following: the size of the network affects the longevity of the endured social contacts within the kin and the friend networks, referring to long-lasted contacts in more extensive kin networks than in smaller ones; furthermore, there is a distinction between kinships and friendships (Roberts & Dunbar, 2011). Palchykov et al. (2013) remark that the frequency of calls does not necessarily refer that it is the contact(s) to whom the most time is spent in communication. Emotional closeness is often related in mobile phone studies to reciprocity (Carron et al., 2016; Puura et al., 2018).

When it comes to the attributes like age, sex, income or others, some patterns are noticeable. For instance, older actors are more prone to long-time contact than younger ones regarding, in both cases, face and non-face interactions (Roberts & Dunbar, 2011). Age and sex are also noteworthy attributes regarding the strength of friendships; they are significant factors for distinctions between the closest and less close friends (Palchykov et al., 2013). The strength of relationships is prone to alter from strong to weaker (Saramäki et al., 2014; cit. (Palchykov et al., 2013) due to the maintenance of newly emerged close relationships. The social brain hypothesis explains the replacement of old strong ties with new contacts, according to which the human evolution of brain capacity limits the number of social ties one has by 150 (Dunbar, 1998). The number of close contacts among them is small, varying from 3 to 5 (Hill & Dunbar, 2003; cit. Palchykov et al., 2013). The increased number of friends affects emotional closeness, which shapes the layered structure of the ego's friendship network, revealed by the mobile phone dataset and refers to the rise of layer size in the structure in parallel with the decrease of emotional closeness (Carron et al., 2016).

In terms of filling the gap between spatial mobility studies and social networks, significant contributions are made using mobile phone data (Puura, 2022). Study based on mobile phone data, such as call detail records (CDR) and mobile phone call graphs (MCG), for extracting spatial and attributive information of calling partners, respectively, revealed a high relationship between the number of calling partners of individual and the extent of their spatial mobility, as well as the diversity in visited administrative units such as a district (Puura et al., 2018). Besides, the spatial dispersion of social networks relates to higher spatial mobility (Puura, 2022).

The characteristics such as sex considerably influence also the patterns of the relationship of an individual's network of calling partners and spatial mobility. On the other hand, attributes such as language and age have a relatively slighter influence (Puura et al., 2018). Sex is also an important attribute when discussing the spatial distribution of networks, referring to the fact that men are more inclined to have spatially more dispersed networks and are spatially more mobile compared to women (Puura et al., 2022).

Social network analysis is also involved in the segregation field, one of the main focus areas in human geography. For instance, the ethnolinguistic composition is one of the attributes that came to the attention of the researchers in studying the relationship between social networks and activity space (Silm et al., 2021). The observations of networks among ethnolinguistic groups, based on the case of majority and minority groups in Tallinn, demonstrated that individuals are prone to more closed networks than open networks. The patterns are remarked for both Estonian-speaking and Russian-speaking ethnolinguistic groups. An interesting point from the study outcomes is that an individual is prone to have a more open network when a higher share of the living area comes from another ethnolinguistic group (Silm et al., 2021).

Further studies of segregation through mobile data and social network analysis is dedicated to investigating segregation based on gender attribute that reveals gender-based disparities (Goel et al., 2021). According to the study outcomes, men are more prone to

gaining information faster than women since they are characterised as being tightly linked. The demonstrated overall findings of the study based on gender, age and language patterns indicate the power of using mobile data and approaches of social network analysis in social sciences (Goel et al., 2021).

Settlement hierarchy is an important facet when discussing social interactions. By focusing on cities, the population density benefits the increase of dyadic communications in number; moreover, the higher population density prompts the increase in the quantity and longevity of phone calls (Büchel & Ehrlich, 2020). Dense urban areas support the complementarity of two modes of interaction, face-to-face communications and calling activities, though this is more significantly exhibited in the localised networks. Additionally, the vast proportion of established phoned contacts are also localised. Concerning cities to their peripheries, it is essential to note that the differences are revealed in calling activity, such as frequency, duration and number of contacts between the city and its periphery. The distance increase encourages decreasing mobile phone-based interactions (Büchel & Ehrlich, 2020).

Using passive mobile positioning data in social sciences has tremendous importance and sometimes is more advantageous than traditional data sources, creating the perspectives for new insights in the studies. For instance, in tourism studies, the advantage of mobile data over typical tourism statistics lies in spatial and temporal precision (Ahas et al., 2008). However, using mobile phone data, particularly in spatial-based studies, is challenging as it requires applying the methodology for detecting meaningful places.

Ahas et al. (2010) have developed a methodology to identify relevant, meaningful places, particularly regularly visited places by the individual, such as home and work, referred to as anchor points. The main parameters used in defining the anchor points are the frequency of calls, daily call activity, and time attributes. The methodology also has limitations, such as having more than one home or work-time anchor point (Ahas et al., 2010). The approach of identifying meaningful locations for spatial analysis is used in many studies (Puura et al., 2018).

3. Data and methodology

3.1. Research area

The study is based on the country of Estonia, located in northeastern Europe, in the Baltic region, the northernmost of the Baltic states and covers an area of 45,339 square kilometres (Encyclopedia Britannica, 2022). The population size is 1,331,824, as present in 2021, from which the share of women is 52.4% and of males - 47.6%. The age composition is the following: the share of the age group under 15 is 16.4%, 63.2% comes to the age group of 15-64, and 20.4% applies to the age group above 65 (Statistics Estonia, 2021).

The ethnolinguistic composition of the country is diverse, though represented by small groups. Estonians make up the vast majority, with 69.1%, followed by Russian, at 23.7%, Ukrainian at 2.1%, and Other groups with 5.1%. The vastly spoken languages are Estonian and Russian (Statistics Estonia, 2021).

The two main patterns characterise the population distribution across the country; First, a higher concentration of the population is shown in the capital and other main cities. Approximately 1/3 (nearly 33%) of the population is present in the capital city, Tallinn, with more than 400 000 inhabitants. Around 17% of the population is concentrated in other big cities, varying between 30,000 and 100,000 inhabitants. The biggest among them, Tartu, is located in the south, balancing the population distribution between the north and south. Other bigger cities are also relatively dispersed in different parts of the country, supporting the balancing of the population distribution. Nearly half of the population is spread in smaller cities, towns and villages. The second pattern is the population density per sq. km. indicating low density in a vast part of settled areas, revealing a high density mostly in big cities.

According to the study purposes, the number of subscriptions of mobile cellular phones per 100 inhabitants is 145, as present in 2020 (Central Intelligence Agency, 2022), ranking first among the European Union countries (The World Bank, 2022).

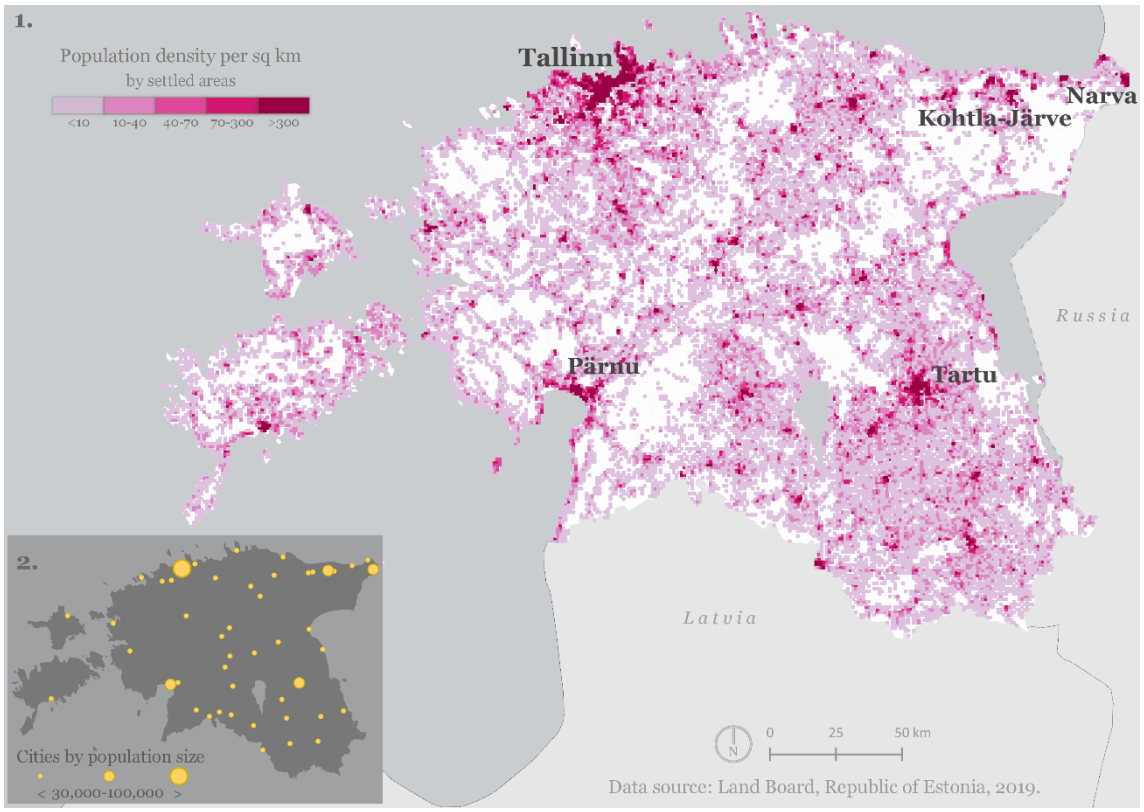


Figure 1. The study area. The population distribution in Estonia. The maps display: 1. Population density per sq. km. by settled areas; 2. The distribution of cities by population size.

3.2. Data

3.2.1. Mobile data

The primary data source in this study is mobile phone data, particularly passive mobile positioning data, provided by one of the biggest mobile operators in Estonia. Passive mobile positioning data is collected mainly by the mobile network cell (indicating the direction of the tower antenna) and is automatically stored in log files of mobile operators that enable anonymous procedure of geographical data aggregation without violating personal information. Finally, serving scientific purposes in various kinds of research (Ahas et al., 2008).

The cellular network consists of base stations with a tower (usually one) and a few directed antennas for a station, forming the cellular network (Ahas et al., 2010). The unique identifications and geographical coordinates are provided for each cell. The noteworthy characteristic of the network cell to be pointed out is the size which is not fixed but instead it is based on switching to the closest antenna. In crowded network circumstances, the phone might switch to any other antenna in the neighbourhood rather than to the nearest one. Also, the density of the coverage of antennas is different in densely and more dispersed inhabited areas (Ahas et al., 2010). These are crucial factors to be considered while discussing the outcomes.

The mobile phone data used in the study refers to the calling activity, meaning both outgoing and incoming calls (Ahas et al., 2008) and is understood as social events (Ahas et al., 2010).

Since the data contains sensitive information, it is stored on the Workstation, managed by the Mobility Lab of the University of Tartu. Access to the data is gained through the mobility lab, regulated by the signed agreement on the use of confidential data of the University of Tartu Mobility lab. All the obligations defined by the agreement are maintained during the study process.

In the study, the raw dataset is used, containing 47,641,197 rows. Each row indicates the call made from the mobile phone between two users, one of which is a call starter and the other a call receiver. Since a call has a direction (starter - receiver), each row refers to incoming and outgoing calls. On the other hand, each row indicates call pair with the representation of two actors of a pair and an interaction through calling activity.

The information stored in the original data includes the details of the base station, the user identifications, call start time and the location of the antenna cell that detected the caller's (call starter) calling signal (see Table 1). Thus, these are attributes for each call pair. The origin-destination matrix is created based on the antenna cell location to analyse the social network of calling pairs with the spatial context; the origin refers to the location of the call starter and the destination to the location of the call receiver. Deriving the origin and destination locations is based on two main steps, described thoroughly in the methodology section: first, detecting home locations; second, identifying call receivers among the column of call starter so the antenna cell location would be known. Completing these steps affected the omission of the observations, reducing the size of the dataset by 86.8% and subtracting only 13.2% for usage in network analysis (see Table 2). The substructed dataset contains call pairs referring to calls between two users, and the frequency of each unique call pair indicates the frequency of calls for a given unique pair.

Table 1. Variables of original dataset.

Original variable name	Meaning
op_pos_usr_id	Caller
op_pos_usr_id_rec	Callee
pos_time	Start time of the call
cell_id	Identification of antenna cell
site_id	Identification of antenna
lon	Coordinate
lat	Coordinate

The study period covers the whole month of February 2018, containing twenty-eight days, and equally representing days of the week. The filtering procedures are applied, affecting

the size of substructed data. Variables are derived from the original dataset (see Table 2), which is described more thoroughly in the methodology unit.

Table 2. Original and final dataset.

Number of rows	> 47,6 m
Year	2018
Month	February
Number of days	28
Size of filtered dataset (rows)	> 6,3 m
Main variables derived from the original datase	Time attributes
	Users' living location coordinates (by antenna)
	Settlement type of users location antenna
	Settlement type of origin-destination links
	Physical distance between call pairs

3.2.2. Additional data

The study uses additional datasets, such as administrative units, cities and settlements, primarily gained from the Land Board of the Republic of Estonia. Most parts of the additional data are used for understanding the context and supporting the visualisations. However, the dataset of settlements has significant importance since it is used for classifying the settlement types of antenna locations.

By Land Board (2022), the settlements are coded according to the following: county, rural municipality, town, city, city without municipal status, city district, small town and village. The given types are classified into two main groups: towns, cities (with and without municipal status), city districts and small towns are defined as urban settlements and villages as rural settlements. Though, since cities differ significantly by population size and population density, the type of urban settlements is divided into two groups: big cities and small cities. The class of big cities is assigned to the cities with a population size of more than 30,000 (see Figure 1): Tallinn, Tartu, Narva, Pärnu and Kohtla-Järve. Additionally, the city of Jõhvi is included in this group due to its geographical vicinity towards Kohtla-Järve. The rest entities of urban settlements are given the class of the small city. Thus, three settlement classes are defined: a big city (urban), a small city (urban) and a rural one.

Finally, the settlement types are assigned to the antenna locations of each row in the dataset using the spatial relationship analysis, enabling the social network analysis based on the settlement types.

3.3. Methods

3.3.1. Preliminary data processing

Since the raw dataset is used in the research, the first steps in the analysis are encountered to deal with the noisy data and subtract the meaningful dataset. Handling the considerable amount of data took an important part and effort in the beginning stage of the study. This chapter thoroughly describes the steps and methods created throughout this phase.

The raw dataset included more than 47,6 million observations. The data science tools were applied to manipulate data, including exploratory data analysis and data wrangling operations.

As stated earlier, the settlements dataset is used additionally to define the settlement types of antennas. A settlement class is assigned to each observation based on the location of the antennae.

A further step incorporates the observations on the data distribution by time attributes. The data covers the whole of February, 28 days. The main insights are as follows: The daily distribution of calls is relatively evenly represented with the distinctive patterns between working days and weekends. The average number of daily calls is 1,701,471. The average number of calls during working days is 1,928,493, while during weekends, it is 1,133,917. Thus, on average, the calling activity during weekends is nearly 33.35 per cent less than during working days.

Regarding the number of daily calls made by each user, the patterns show values under 100 for most users, though outliers and extreme cases are also presented. The steps for filtering out the outliers are described later.

The distribution of weekdays in the dataset is also examined to avoid biased outcomes. Examination revealed no gap in by weekdays during the four weeks.

The next step involves analysing the diurnal rhythm of the calling activities, showing the peak of calling activities between 10 AM and 5 PM with a steady increase and decrease before and after peak time, respectively.

The diurnal rhythm by weekdays reveals the differences among weekdays, referring to a lower hourly peak of calling activity on weekends compared to working days. The first few working days pattern similarities, while from Thursday when the peak of activity is more distinguished, making the curved shape by Friday.

Examining the data's various variables revealed the requirement for applying filtering tools to remove noisy and redundant observations. The first step involved aggregating daily calls by call starter, and calculating daily call mean and median numbers. The given median number of 3 and the mean of 5.1 indicates that the outliers affect the higher mean

number. For a more intuitive understanding of the distribution, the percentiles are calculated, showing the following result:

Table 3. The percentiles of daily call activity by users (before filtering).

Percentile (%)	25	50	75	90	95	99	99.5	99.9	99.95	99.99
Value	1	3	6	11	16	32	43	79	110	207

First, the observations falling over 207 are extracted, and the percentiles are calculated again, exhibiting the following: 98.9 per cent of observations fall under 30. At the same time, the histogram depicts the density of daily call activities. It visually reveals the long tail of density on the bottom that starts from 30 and has a long continuous pattern from 70. Thus, it is logical to cut the density tail depicted on the histogram starting from 30.

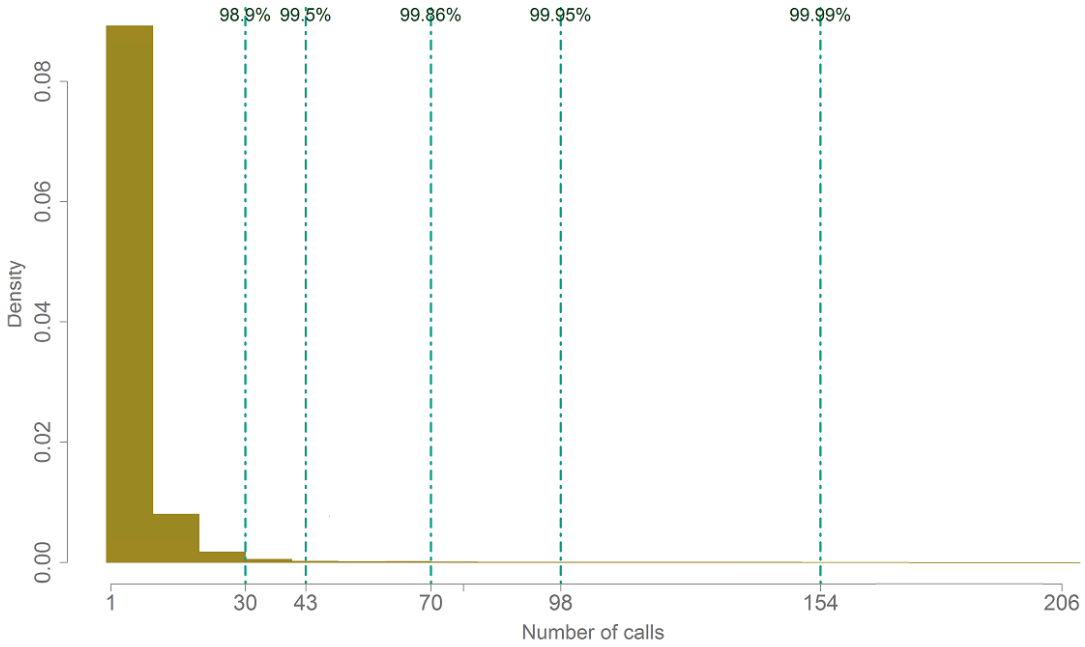


Figure 2. Histogram displaying the distribution of daily call activities by users.

At this stage of filtering the data, the dataset is reduced by 12%, considered noisy observations. However, further steps are still required to complete the filtering.

The next step applies to filtering the data based on diurnal rhythm analysis. The distribution of observations by daytime reveals the rhythm of call activities, though the interpretation of the time variable affects the distribution. In the dataset, three variables represent the hour extracted from the time variable from the original dataset. The first variable that depicts hours is called "hour" and is extracted from the time variable directly. The second variable is "hour_rounded" and is extracted by considering minutes as well, so if minutes in the time is over 30, the hour is rounded to the maximum, otherwise to the minimum; the third extracted variable is "hour_min" that is a combination of an hour and

minute (hour + minute / 60). Using "hour" or "hour_rounded" variables shows biased outcomes for the density distribution depicted on the histogram, especially from midnight till morning hours.

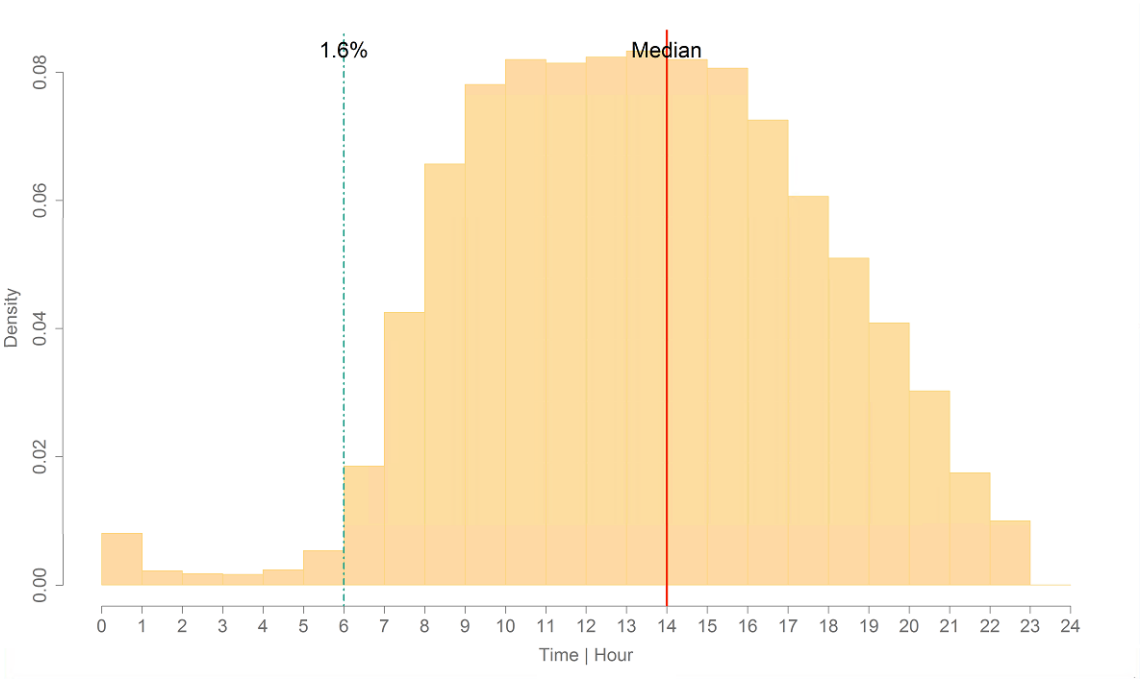


Figure 3. Histogram of daily call activities by hours (with time variable: Hour).

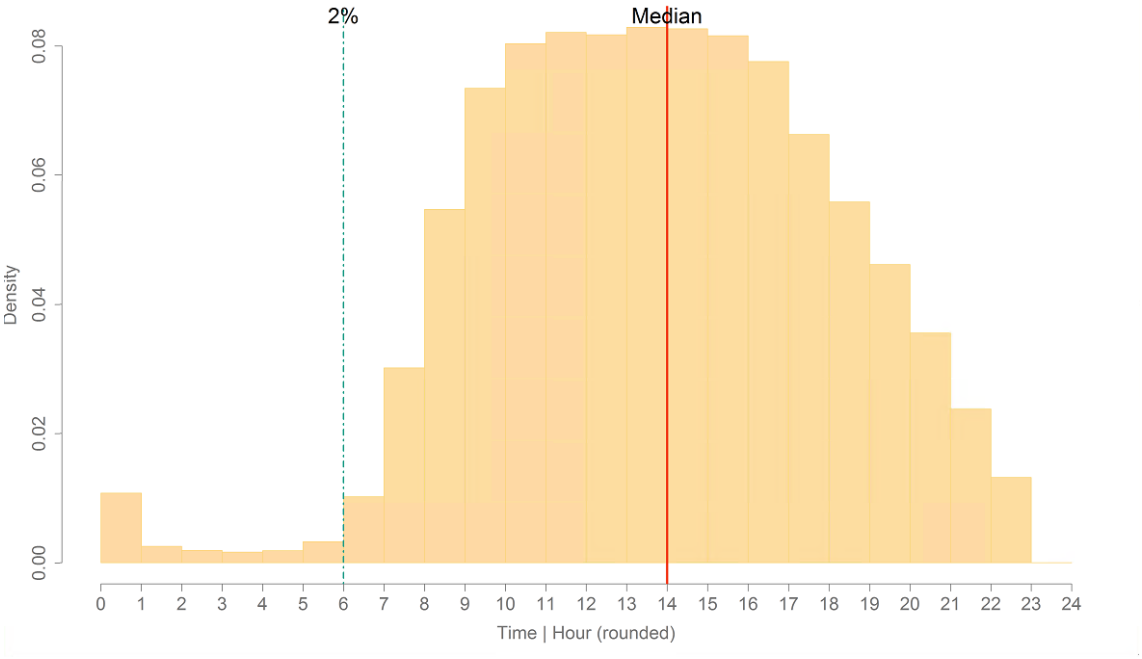


Figure 4. Histogram of daily call activities by hours (with time variable: Hour (rounded)).

On histograms (see Figures 3-4), it is visually evident that the density of hourly call activity decreases drastically between 23 and 24 and increases significantly between midnight and 1 AM, which indicates abnormal distribution. Thus, it is more reasonable to use the combination of an hour and minute to depict distribution more continuously. The histogram where the density of calling activity is displayed according to more precise time (hour and minute) completes the gap between 23 and 24 and reveals the gradual decrease from midnight. The period from midnight to 6 AM might be an outlier since it could be believed that calls made during the nighttime might not reflect peoples' social networks, and these observations are excluded (see Figure 5). This assumption is based on the general perception of the diurnal rhythm on an everyday basis. It considers the circumstances, such as interactions with actors living in different time zones, calls for emergency services or urgent situations, etc., as a possible explanation of nighttime calling activity.

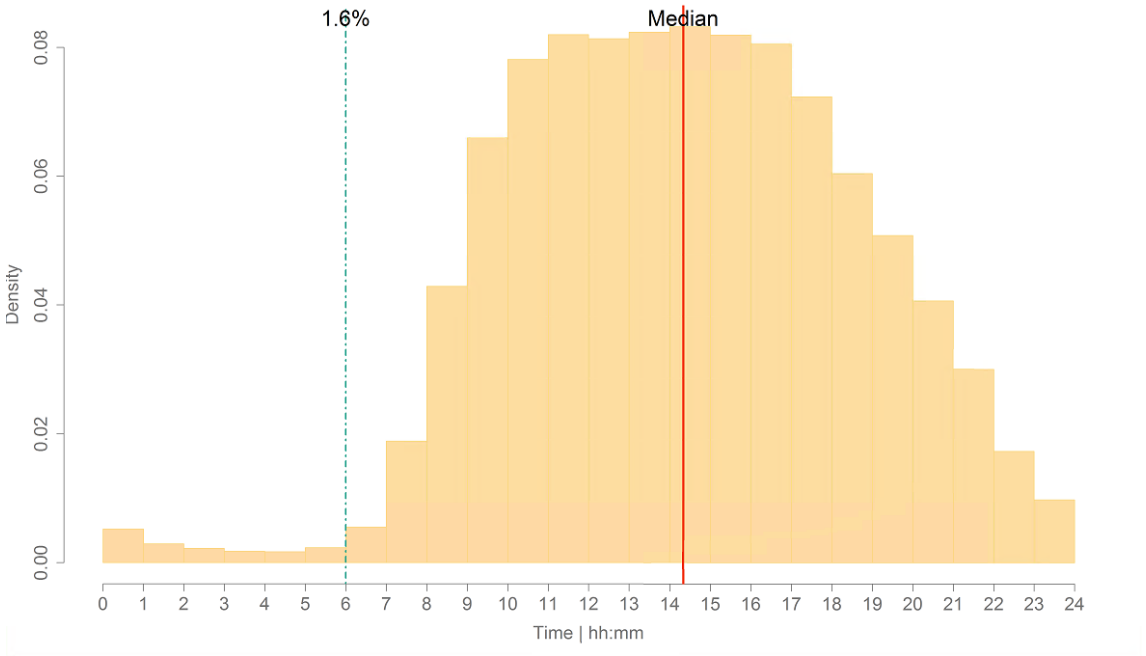


Figure 5. Histogram of daily call activities by hours (with time variable: Hour and Minute).

Visual observation of calling activities aggregated by hours refers to a relatively normal distribution with a bell-shaped pattern (see Figure 6). The number of calls increases gradually between 7 and 10 AM. Calling activity is relatively stable from 10 AM to 5 PM, with a slight concave pattern by 1 PM. After 5 PM till midnight, a gradual decrease is depicted. Thus, before 10 AM and after 5 PM, two opposite tendencies occur, increase and decrease of calling activities, both with gradual patterns.

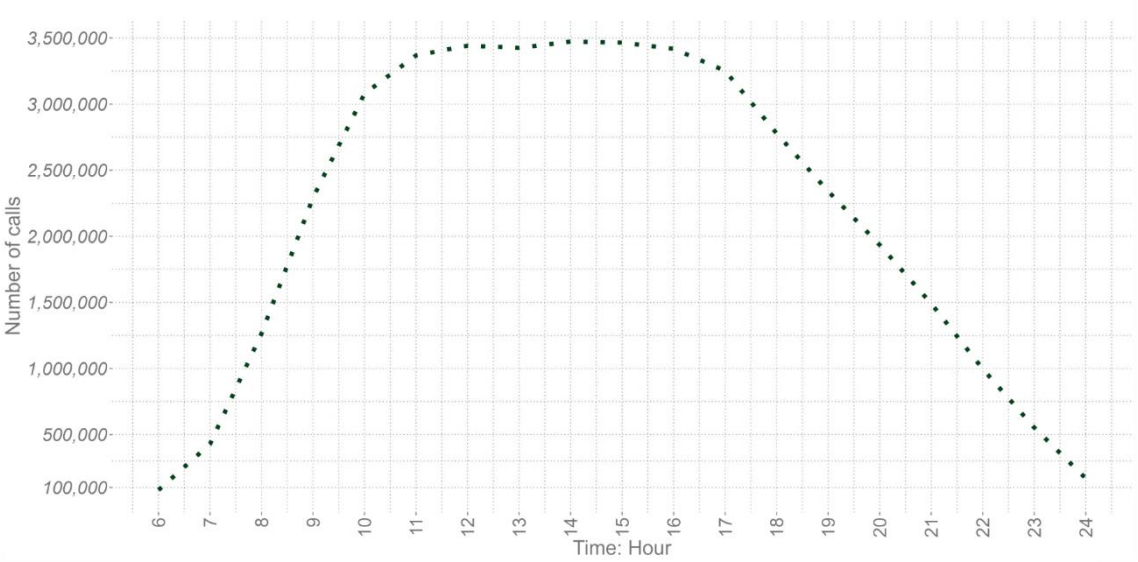


Figure 6. Hourly distribution of calling activities on the filtered dataset.

Patterns of the diurnal rhythm of calling activities by weekdays differ during working days and weekends regarding the number of calls. However, the distribution characteristics by hours are relatively similar with slight differences (see Figure 7).

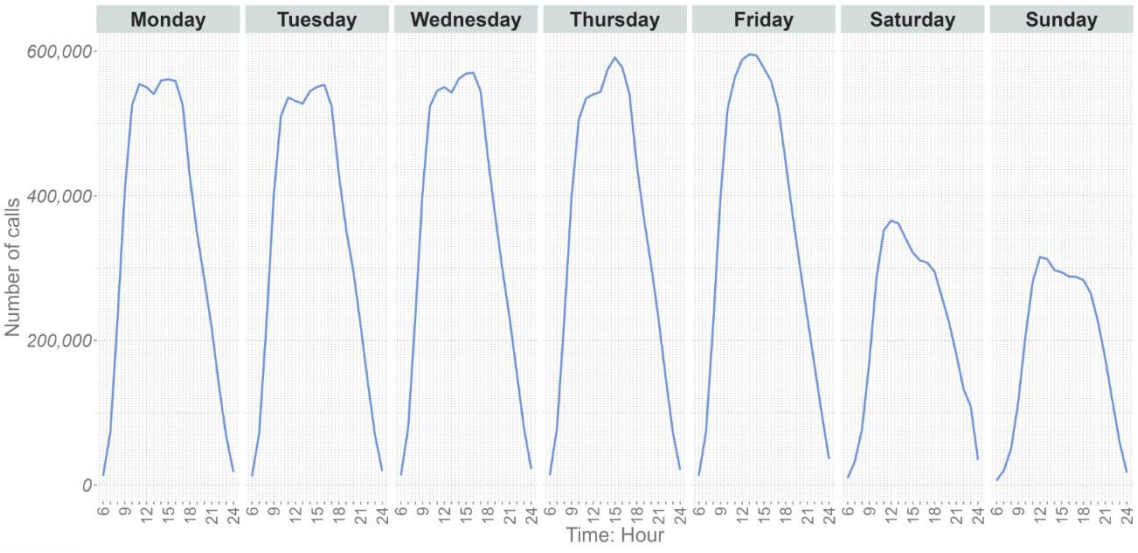


Figure 7. Hourly distribution of calling activities by weekdays on the filtered dataset.

Thus, this chapter demonstrated the steps required for filtering from noisy observations and preparing the dataset for further actions.

3.3.2. Detection of home locations

The primary purpose of this study is to conduct an analysis based on the locations of call pairs. Though the dataset contains the antenna location for only the call starter. The developed methodology to derive the spatial context based on call pairs for origin-destination locations is the following: First, to detect meaningful locations such as home

and consider this location as the user's living area for all the observations the user appears. Second, to identify the call receivers among the list of call receivers and assign the living antenna location. The limitations of the methodology are discussed in later chapters. Home anchors are detected by previous studies as well (Ahas et al., 2010), based on the dataset corresponding to 2008. However, the current concepts of teleworking and work-from-anywhere have prompted remote working, part-time working, and freelancing, making it even more challenging to define and standardise working hours and define the home and work locations due (Choudhury et al., 2021; (Türkeș & Vuță, 2022). Therefore, the methodology to detect home locations is developed independently from the previous studies.

The first step of the methodology relies on the diurnal rhythm of calling activities by using the hourly rounded time variable.

Boxplot displays (see Figure 8) that calling activities during the day occur mainly between 11 AM and 5 PM, making this part of the day the most active and possibly referring to activities outside, such as work and study purposes. The period before and after this time frame indicates relatively passive rhythms.

Figure 8. Diurnal rhythm of calling activities.

Dividing a diurnal rhythm into two parts, home and working hour periods, is rough and excludes the additional activities that occur between or during these periods. Due to data limitations, detecting the location and time of other activities is challenging. Thus, identifying meaningful locations is limited to primary types like home and work.

From the dataset, the calls made between 11 AM and 5 PM are filtered out, and only calls made before and after this time frame are maintained. Filteration kept only 42.2% of the observations from the previously filtered dataset. The removal of these observations is made to detect home locations and does not refer to excluding all the calls between specified hours from the finally filtered dataset.

The filtered dataset locations are considered the potential home places, though further steps are required to assure better precision of the methodology. The next phase is based on the frequency of the days the user has made calls from identical antennas, assuming that the antenna that repeatedly detects the user's call signal over the days is most probably the home location.

The display of the aggregated data by the call starter, antenna and day exhibits each user's frequency of calling activity from the repeated antenna (see Figure 9).

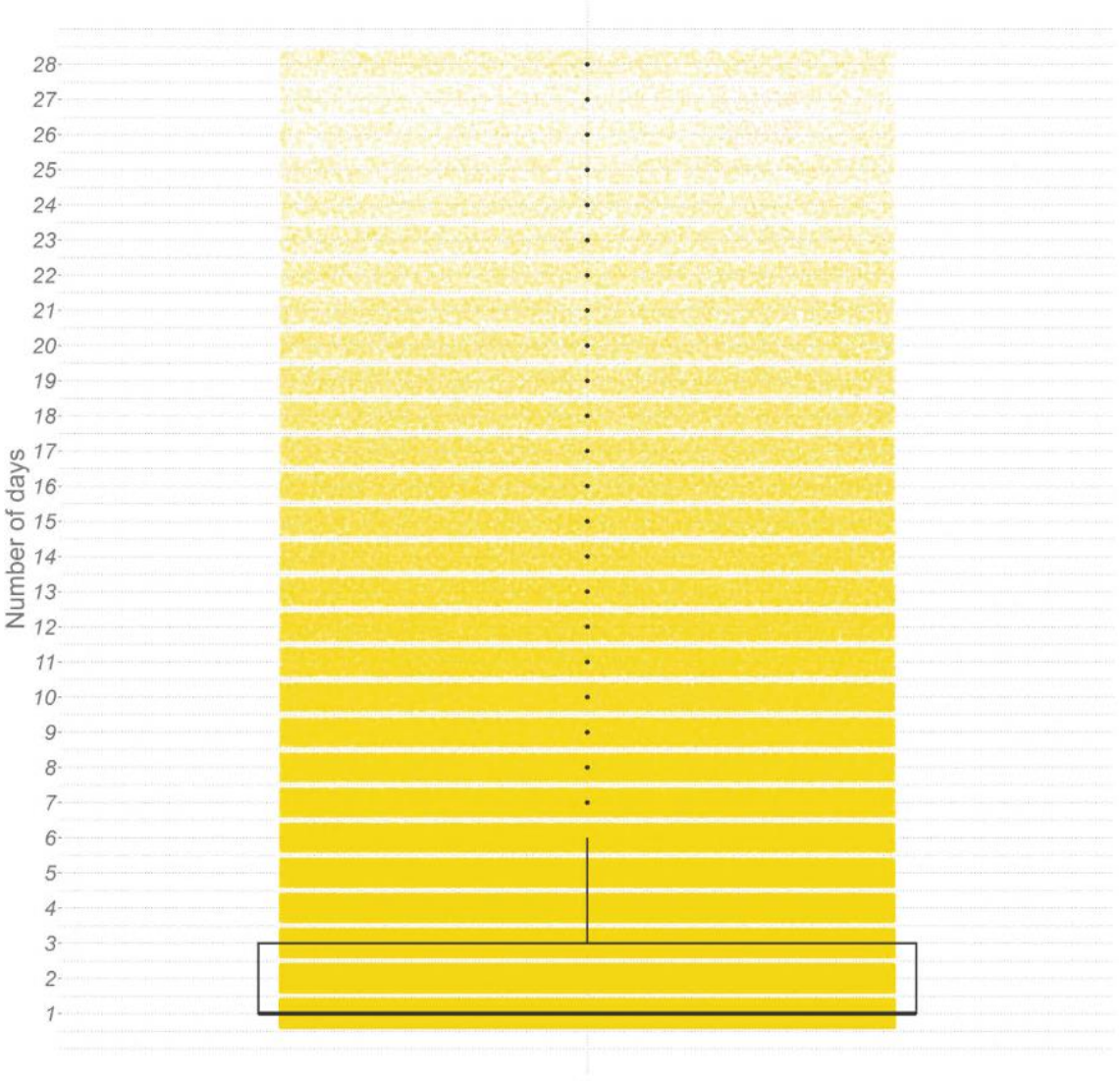


Figure 9. Frequency of active calling days by callers.

For the vast majority, the frequency of calls from the same antenna is made only in one day, with no repeated activities from the same antenna on other days. The median is 1, while the third quartile is 3. For 50% of users, the calls from the same antenna are made only for a day, while for 75%, calls are made from the same antenna repeatedly for three days.

It is challenging to distinguish random calls, results of errors and correctly detected calls from each other, especially when the frequency of calling activity days from the same

antenna location is low. Thus, the cases when calls are made by callers from the same antenna no more than three times are omitted, and only cases with more than three frequencies are kept.

The next step is to calculate the frequency of antennas by each caller. The boxplot below shows the number of antennas counted for each user (call starter). The median frequency of antennas is 1, which means that for 50% of the observations, the number of meaningful places is one. While for 75%, it is up to 2, and for 90%, it is up to 3.

Table 4. The percentiles of frequency of antennas by callers.

Percentile (%)	25	50	75	90	95	99	99.5	99.9	99.95	99.99
Value	1	1	2	3	4	6	7	9	10	12

Boxplot displays that the maximum frequency is three, and higher numbers are outliers. When talking about meaningful places, it is plausible that a person has more than three meaningful places, including locations for additional activities besides home and work-related. Though, about home locations, it is not likely for a person to have more than three home locations. Since the focus is on detecting home locations, the users (callers) with more than three frequent antennas are omitted (nearly 8%).

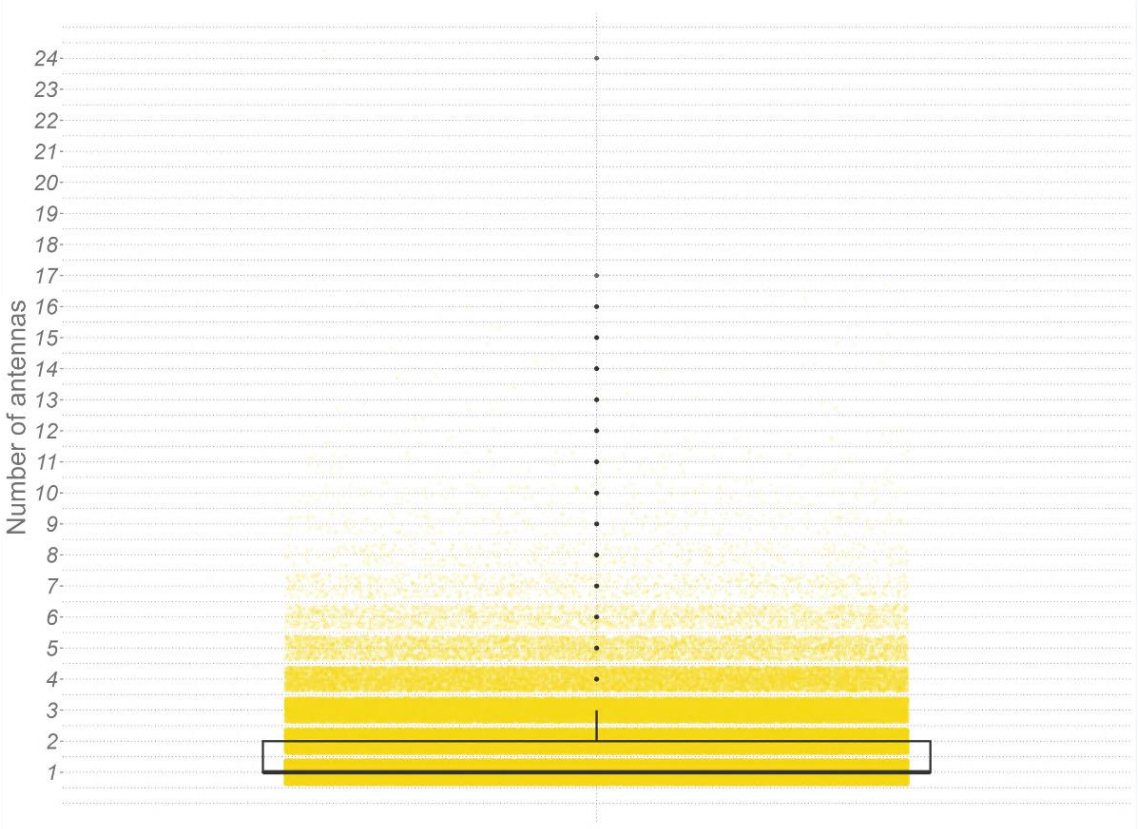


Figure 10. The frequency of antennas by callers.

Considering home as a place where people make the most calls outside active hours, the antenna locations with the most calls are kept. In cases of an equal amount of calls when a person has two or three locations, the antenna is selected randomly.

Thus, the filtered dataset at this stage incorporates the observations with detected home locations for the call starter but not for the call receiver. It is assumed that call receivers are also callers, so it is possible to identify call receivers among the call starters. First, the common users appearing among both variables of callers and callees are identified. After, the callers are assigned the antenna locations based on the column of the caller. Since locations for both users are assigned, the dataset is filtered based on common user identifications.

Finally, the filtered dataset has origin-destination characteristics, origin representing the antenna location of the call starter (caller) and the destination - the antenna location of the call receiver (callee).

The origin-destination characteristics of the dataset enable measuring the straight physical distance between the antenna locations of call pairs.

The dataset is designed for implementation in social network analysis. The final dataset contains 6,309,152 rows, 321 162 unique callers, 320 497 unique callees, and 1 667 294 unique calling pairs.

3.3.3. Network analysis

The "igraph" package is used in R to create the graph object from the dataset of mobile calls. The graph is defined as directed by setting the "True" logical value for the direction parameter. The attributes are assigned for both vertices/nodes and edges/links, so in later steps, network analysis and visualisation are attribute-based. Two main functions used throughout network analysis are described below.

Network centrality

For estimating the network centrality, the calculation is made on vertices in the social network of call pairs referring to actors by applying the degree function. The degree of a vertex indicates the number of adjacent edges for a vertex.

The author of the function degree {igraph} is Gabor Csardi, and the implementation of the calculation is accomplished according to R Documentation (Igraph, 2022). As a first argument to analyse, the graph object is assigned. The second defined argument considers the ids of vertices on which the degree is calculated. In this case, all the vertices are put into the evaluation. The following argument referred to in the calculation is related to the mode of edge direction and is important for interpreting the outcomes depending on the given parameters. As stated in the theory part, node degree has three different measurements, in-degree, out-degree and all-degree, defined by the mode argument as "in", "out", and "all" (or "total"), respectively. For analysing network centrality, each mode parameter is applied and calculated separately. The estimated numeric values are

given for each vertex of the graph. Higher values indicate a more central location in the network than vertices with lower values.

The degree distribution is also estimated by applying the additional argument in the calculation. Besides the mentioned arguments, cumulative degree distribution is calculated additionally.

For visualising the spatial patterns of degree estimation, Gephi software is used (Gephi, 2022). The "GeoLayout" plugin is operated, enabling displaying the graph based on geocoded attributes such as spatial coordinates and standard projections. The plugin is developed by Alexis Jacomy. For displaying the graph spatially, Mercator projection is specified.

The average weighted degree is also calculated in Gephi. The computation considers the weight (frequency of links presented in the network) when calculating the degree and provides the average weighted degree values based on each mode.

Strong and weak ties

Calculations of edge weight and edge betweenness are applied to estimate the strength of ties and detect strong, weak and vulnerable social links. Edge weight indicates the frequency of links in the network. In this study, the weight refers to the intensity of interactions of pairs by the frequency of completed calls. The edge weight is derived from the simple aggregation calculation or calculated when creating the graph object. The edge betweenness calculation is based on a more comprehensive function.

The edge betweenness indicates the number of shortest paths (geodesics) going through an edge (Igraph, 2022). The credit for the authorship of the function belongs to Gabor Csardi. The function is based on the references such as Centrality in Social Networks (Freeman, 1978b) and A faster Algorithm for Betweenness Centrality (Brandes, 2001).

The arguments for estimating edge betweenness are defined according to R Documentation. The first argument indicates the graph object that is to be analysed. The graph is defined as directed by assigning a "True" logical parameter for the direction argument. The weight parameter is also set, so the weighted shortest paths, interpreted as distances, are considered. The cutoff argument is also defined, which refers to the maximum path length to be considered while calculating the betweenness. Setting the argument to an unlimited parameter is time-consuming and requires a computational capacity to calculate the edge betweenness for a large graph. The cutoff argument is set according to the network centrality analysis. Node degree estimations have revealed that the median number of adjacent edges for nodes is four based on both in-degree and out-degree mode calculations. Therefore, the cutoff argument is set at four. The calculation gives the edge betweenness scores for each edge of the graph.

3.3.4. Tools and software

A big part of the analysis is conducted in R using the RStudio environment, specifically, the R version of 4.2.1. and RStudio version of 2022.07.02 Build 576. R is used for

analysis as well as for visualising purposes. The main libraries used are the following, though not limited to:

- "tidyverse" for statistical computations
- "lubridate" for extraction of time attributes
- "stringr" for manipulations on text
- "ggplot2" for visualisation
- "ggpubr" for publication plots
- "geosphere" for spatial manipulations
- "igraph" for network analysis
- "rmarkdown" for generating the rendered output of the analysis

For the network analysis, additionally to “igraph” package in R, the open graph viz platform, Gephi, is used with version 0.9.7., for statistical calculations based on built functions and visualising.

For visualising purposes, QGIS is as well used with version 3.10.0.

In the final version of the thesis, open access to the script will be provided through the generated document from RMarkdown and shared on the GitHub platform without revealing the dataset's content due to personal privacy.

4. Results

4.1. Exploratory analysis

The size of the network of mobile call pairs by the parameters of network size is following:

- The network consists of more than 332,7 thousand unique nodes, here actors, representing nearly 25% of the population of Estonia.
- The number of links, representing the total calling activities between calling pairs, is more than 6,3 million.
- The number of unique links, designating the number of calling pairs, is more than 1,7 million.

The type of the network is directed due to its characteristics of call activity involving, on the one hand, initiating a call and, on the other hand, receiving a call. Therefore, the call activity has two actors, the call starter (caller) and the call receiver (callee), in graph analysis indicated as the source and the target, respectively. The numbers of unique sources and targets in the given network differ slightly. The network consists of both reciprocal and non-reciprocal relationships of call partners.

After analysing the network size, the parameter of edge weight is evaluated, which refers to the frequency of edges/links between two particular nodes. The edge weight in the mobile call dataset indicates unique call pairs' frequency of call activities. The results reveal that the minimum call frequency is one, the maximum is 532, the median is 2, and the average number equals 3.8 (see Table 5). The share of call pairs with high calling activities is relatively small.

Table 5. The weight of edges (call pairs) – call frequency.

Minimum		1
Percentiles	25% (Q1)	1
	50%	2
	75% (Q3)	4
	90%	8
	95%	14
	99%	33
	99.50%	45
	99.90%	82
	99.95%	101
	99.99%	156
Maximum		532
Average		3.78

Visual analysis of mobile call partners' social network in Gephi software using ForceAtlas2 layout displays the characteristics of the network (see Figure 11). The size of the network resulted in a not extremely precise outcome. However, essential particulars can be derived from the graph's visual representation where the actors' placement is displayed. First, the core part of the middle of the graph displays several grouped parts, indicating the actors' closeness within the groups. Second, the graph has a surrounding circle formed by the nodes and pairs disconnected from the core part of the graph. Besides, weakly connected nodes are represented around the core part of the graph (see Figure 11).

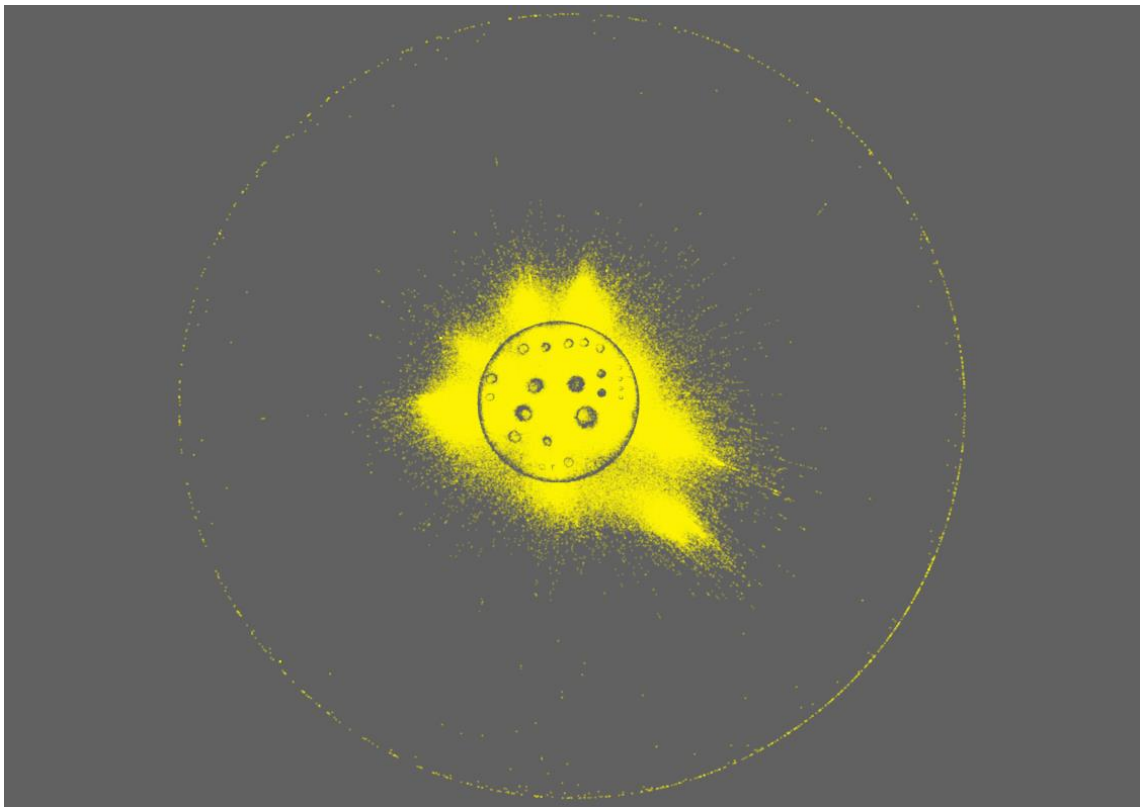


Figure 11. The social network of call pairs. The network depicts how close the actors are to each other based on the settlement type. The network is created in Gephi by using Yifan Hu Multilevel layout.

Examination of the node representation in a graph by the settlement types shows the different patterns (see Figure 12). Few noticeable groups represent nodes with settlement types of big urban cities that are more closely connected. Nodes with the settlement types of small urban cities and rural form more noticeable groups that might be interpreted as actors living in small cities and rural areas are relatively more connected compared to settlements of big cities and rural ones. Some groups of connected nodes with all the settlement types form the group in the middle of the graph.

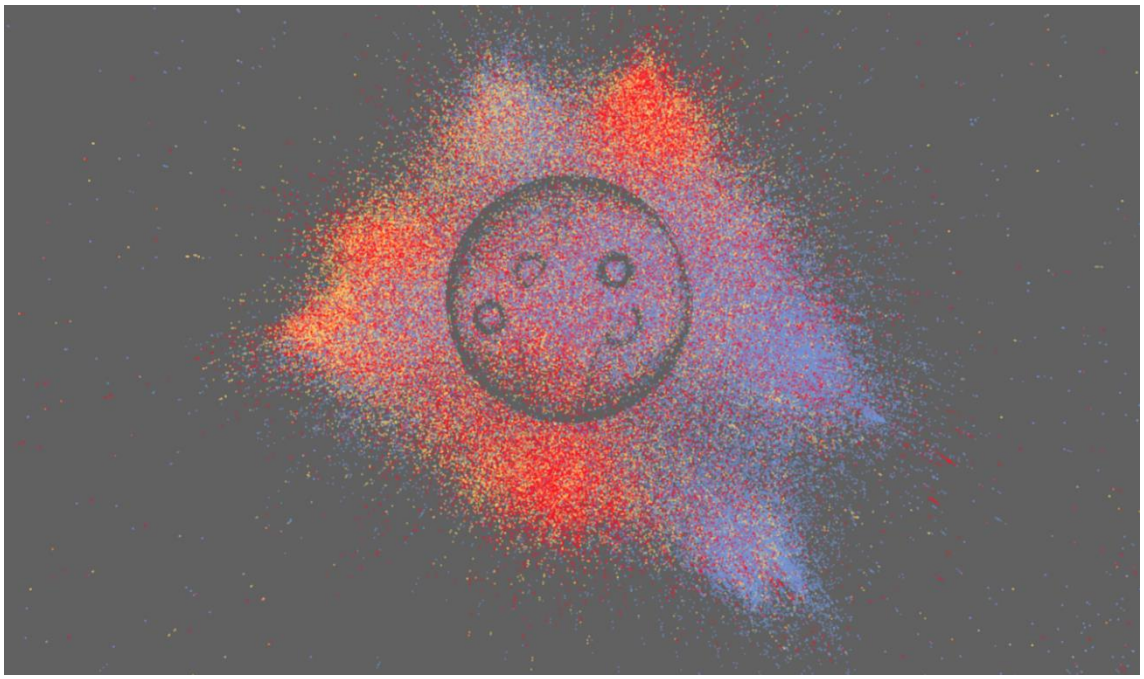


Figure 12. The social network of call pairs. The network depicts how close the actors are to each other based on the settlement type. The colours depict the settlement type of actor: blue referring to big city, red – small city and yellow to rural settlements. The network is created in Gephi by using Yifan Hu Multilevel layout.

The visual display of edges/links by the settlement types reveals that the ties are created mainly between the same settlement types (see Figure 13). The numbers also support the visual examination (see Figure 14). The highest share, 30.24%, of the represented links in the network comes to the ties for which both source (caller) and target (callee) is from the big urban cities, followed by ties with source and target from small urban cities (18.84%) and links with source and targets from rural settlements (13.33%). Thus, 62.41% of the links are created by actors from the same settlement types, including those from the same settlements. The rest of the share corresponds to the links created by the different types of settlements which are more or less equally represented. Though ties between big and small cities are more represented, followed by links between small cities and rural settlements, and finally, a minor share comes from ties between big cities and rural settlements.

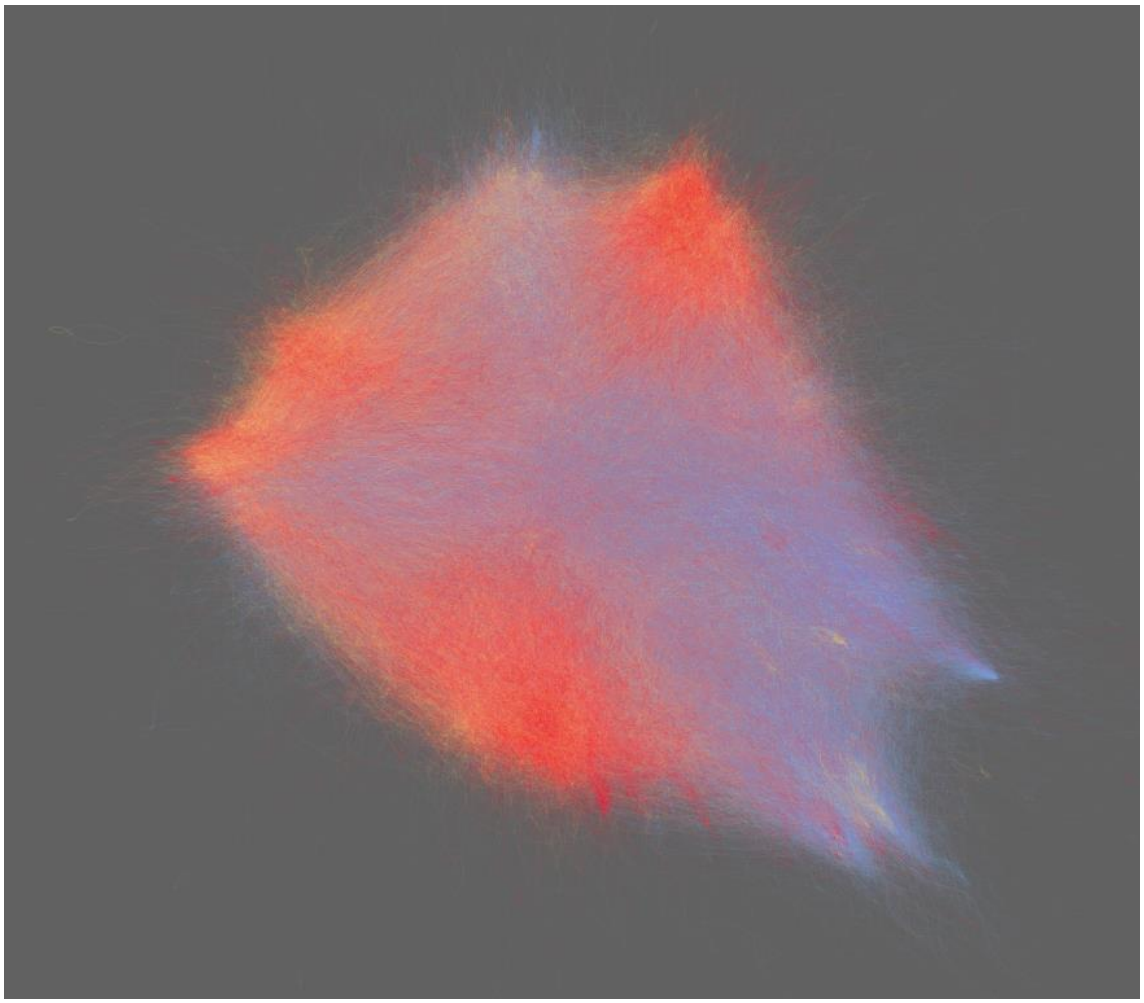


Figure 13. The social network of call pairs. The network depicts how close the actors are to each other based on the settlement type. The colours depict the the different pairs of settlement types of links. The network is created in Gephi by using Yifan Hu Multilevel layout.

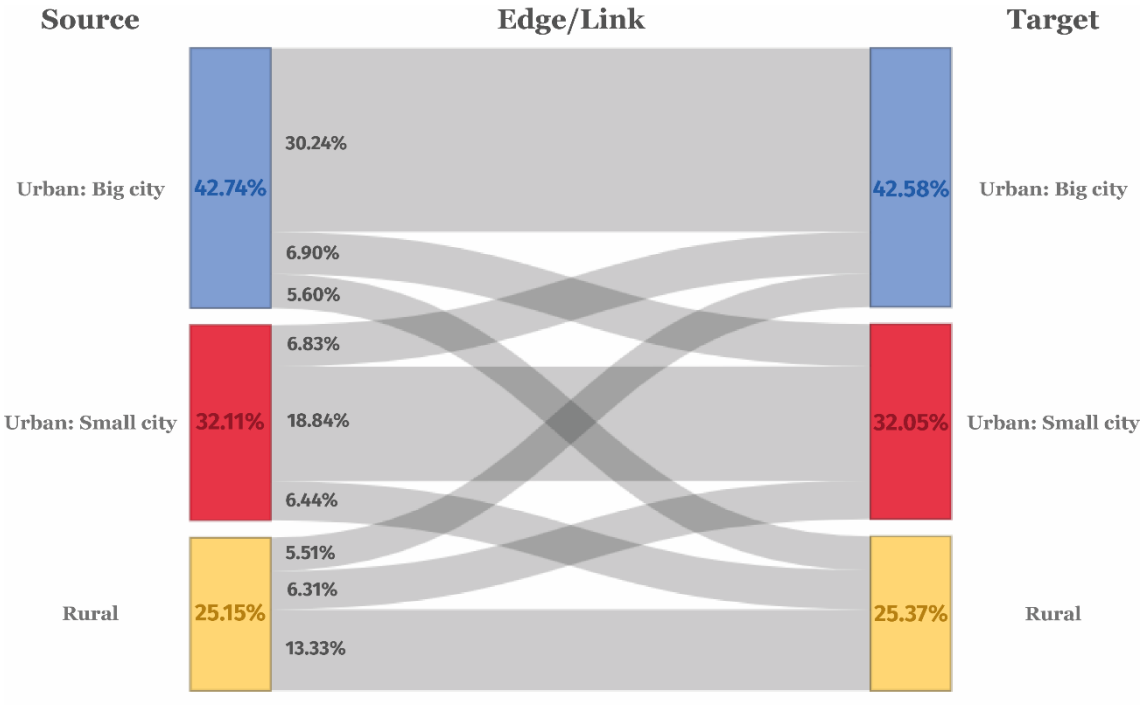


Figure 14. The social links created of calling partners.

The spatial distribution of the mobile call network demonstrates the most heightened concentration of call pairs and calling activities in Tallinn and Tartu, Estonia's most populous cities, followed by Pärnu and other cities (see Figure 15). Visual examination of the spatial distribution shows the spatial extent of concentration of dispersion of call pairs. In the case of Tallinn and Tartu, the concentration is spatially extended significantly, involving the settlements located around and visually forming the grouping pattern. Similar patterns, though with a smaller scale and extent, also apply to other settlements.

Besides concentrating and grouping the calling activity around particular settlements, the relationship patterns between the far-flung settlements are also expressed. The tightest far-flung ties of calling partners are created between Tallinn and Tartu and between Tallinn and Pärnu. Far-flung dyadic ties are represented throughout the county; however, particular settlements are more connected regarding social links (see Figure 15).

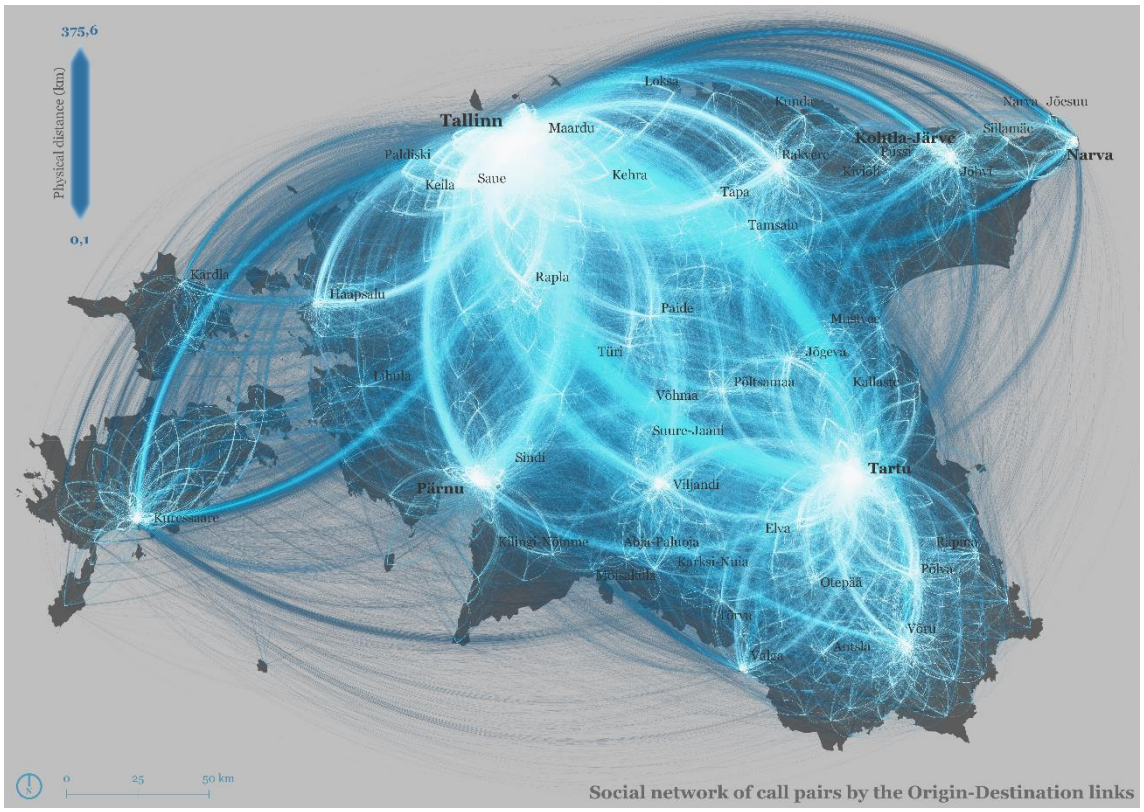


Figure 15. The network of call pairs. The map displays the Origin-Destination matrix of calling partners based on antenna locations.

The settlement types establish the hierarchical relationship, revealed by analysing the spatial distribution of ties by settlement types (see Figure 16). The most robust dyadic relationships are established between big cities, predominantly between Tallinn and Tartu and Tallinn and Pärnu, followed by pairs of Tartu and Pärnu, Tallinn and Narva, and Tallinn and Kohtla-Järve area. The dyadic links are relatively weakly represented between other pairs of cities. Visual examination demonstrates the apparent pattern of reciprocal relationships between all pairs of big cities, meaning that both actors of settlement pairs commence calling activities. Furthermore, the reciprocal links made by each side of the settlement pairs are equally represented.

Relationships between pairs of small cities also have reciprocal nature; however, reciprocity is not shown for all dyads of settlements. The higher concentration of links between small cities is displayed around Tallinn. The latter also applies to the ties between rural areas, though additionally, high concentration is provided in the southern part of Estonia, around cities. Reciprocity is not a vital characteristic in the case of dyads based on rural settlements.

The remarkable pattern in the hierarchy of settlement types lies in the differences in relationships between different types. Representation of established links is more significant between big and small cities, followed by pair of big cities and rural settlements and then by small cities and rural settlements, compared to the pairs of the

same types. The pattern of reciprocity is also displayed for the big city-small city and big city-rural pairs.

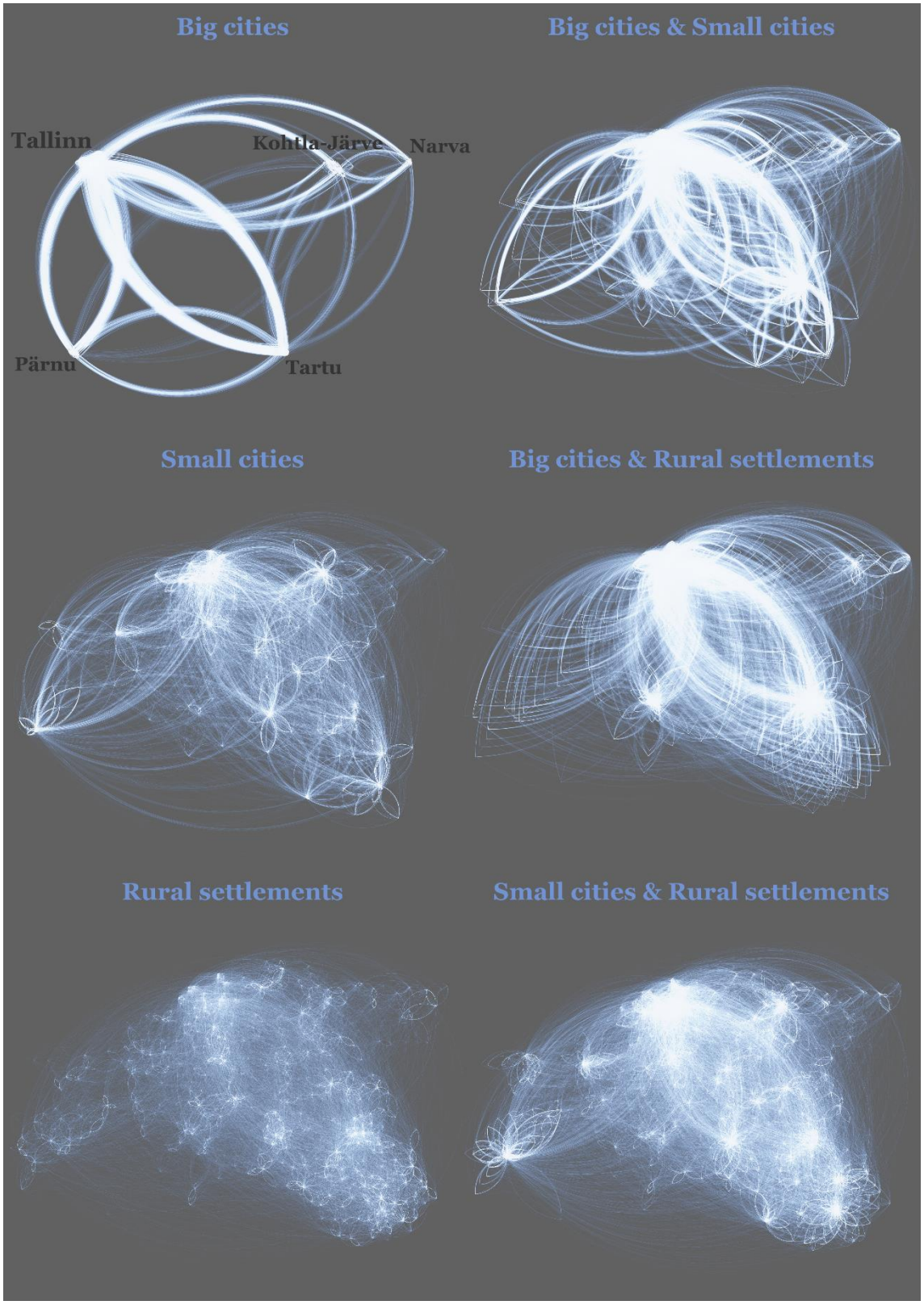


Figure 16. The spatial distribution of social links by the pairs of settlement types.

Note that the map displays the unique links between settlement pairs and does not depict the frequency of calls made by call pairs. Thus, a high representation of links between particular settlements is not directly interpreted as social closeness.

Besides the spatial context, the temporal information is derived from the given mobile dataset. Temporal analysis on weekdays reveals the patterns in different calling behaviour by the pairs of settlement types (see Figure 17). The calls from big cities to other big cities and other types of settlements have similar patterns, with a small peak in the middle of the week and an increasing trend by the end of the week, while other pairs of settlements show different patterns. Analysing the calling activity separately within the same settlement and between different settlements resulted in more compelling outcomes. The trends of calling activities within the same settlements differ from trends between different ones.

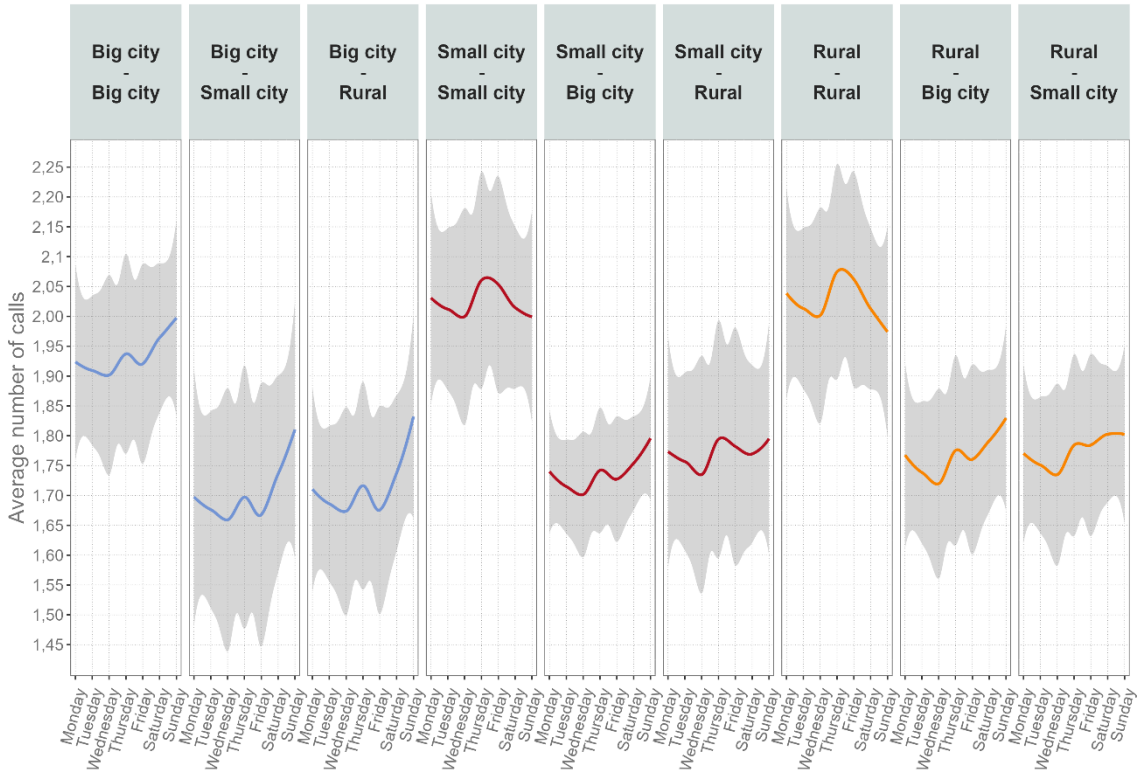


Figure 17. Calling activity by weekdays and the types of the settlement pairs.

4.2. Network centrality analysis

Network centrality is analysed by estimating node degree with three different modes, in-degree, out-degree and all-degree, resulting in values depending on the direction of calling activity - incoming, outgoing and all (both) calls, respectively. Calculating the degree distribution by the degree mode, considering cumulative frequency, reveals that the values are vastly distributed between 1 and 10, referring to the number of linked actors for each node (see Figure 18). The tendency of distributing values less than 100 is the same for both in-degree and out-degree modes. For values higher than 100, out-degree

values are not presented, and in-degree values are delivered though the distribution is relatively slight.

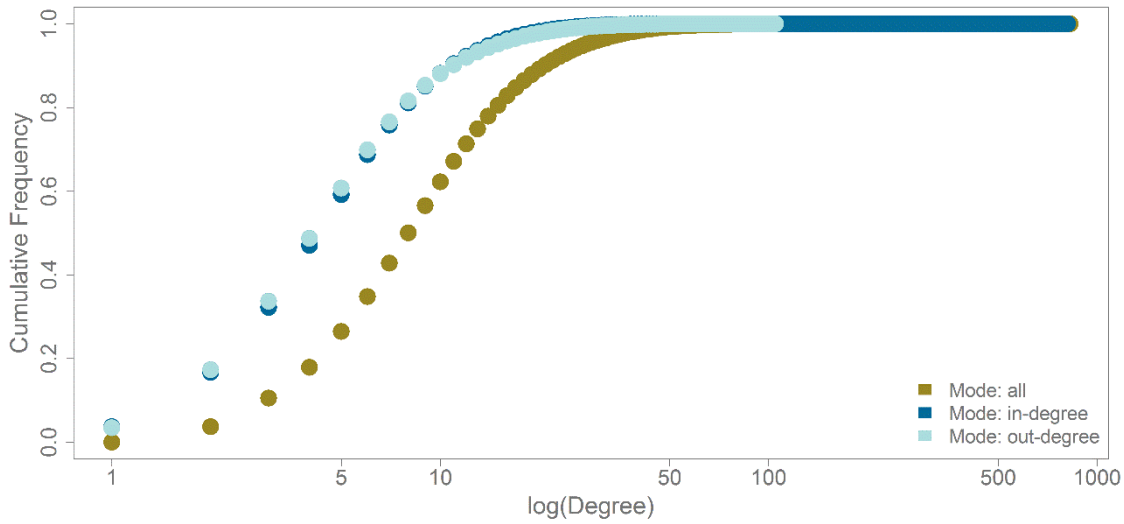


Figure 18. Degree distribution with cumulative frequency.

The outcomes of node degree estimation by applying the degree function are displayed on a histogram showing the density of values by the settlement types of nodes. The logarithm function (log()) is applied to display the values on a histogram. The distribution of values calculated with the mode parameter of all-degree shows that the lower quartile of distributing values is similar by all types of node settlements, presented at 4 (see Figure 19). The median number is slightly different, at 7 for nodes based on big cities and 8 for small cities and rural settlements. While the upper quartile is the same and average numbers are almost similar by the settlement types of nodes, the maximum values of node degree differ significantly. The highest number of degrees are shown for nodes based on big cities, making up approximately 37% and 42% difference compared to small cities and rural settlements, respectively.

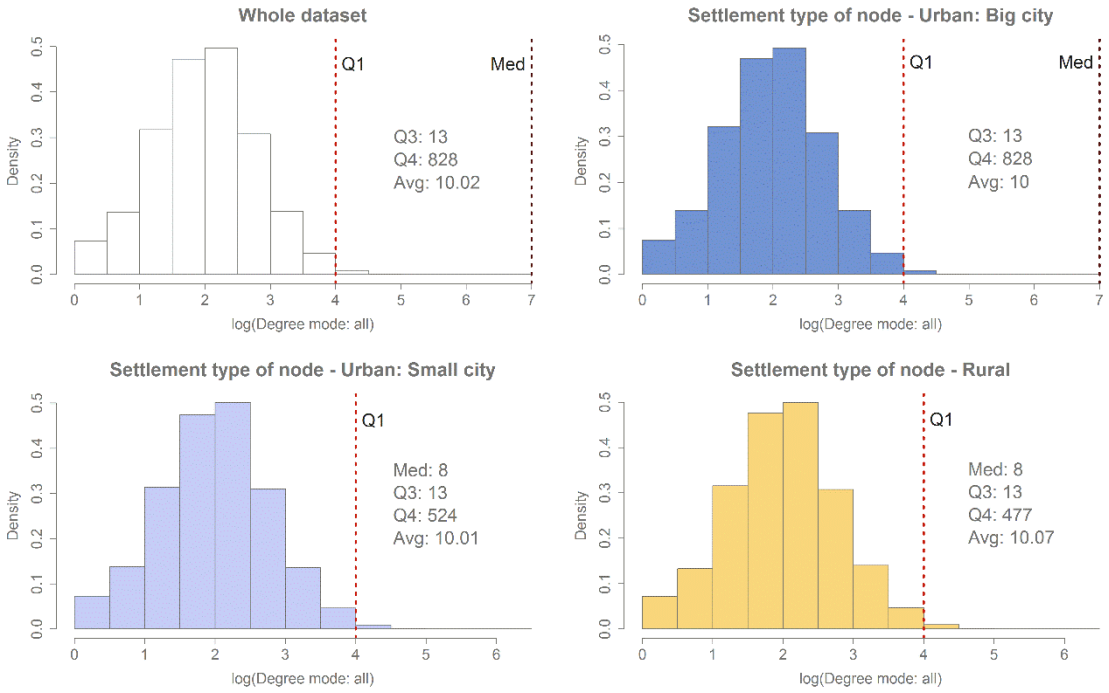


Figure 19. Histograms displaying the node degree by the settlement pairs of links. Mode: All-degree.

The lower quartile, upper quartile and median numbers are similarly presented for all node groups based on settlement types in both cases, in-degree (see Figure 20) and out-degree (see Figure 21) mode. The average number of degrees is also approximately equivalent for both modes and all groups of settlement types. The differences by settlement types and modes occur in maximum degree values. Estimation based on in-degree mode reveals a similar trend of displaying the maximum values as the all-degree mode-based estimation - the maximum value is shown for a group of big cities, making approximately 35% and 42% differences with small cities and rural settlements, respectively. The maximum values of out-degree are significantly low compared to in-degree values. Moreover, the trend contrasts with the in-degree mode indicating the highest number for rural settlements, though the difference is not extreme.

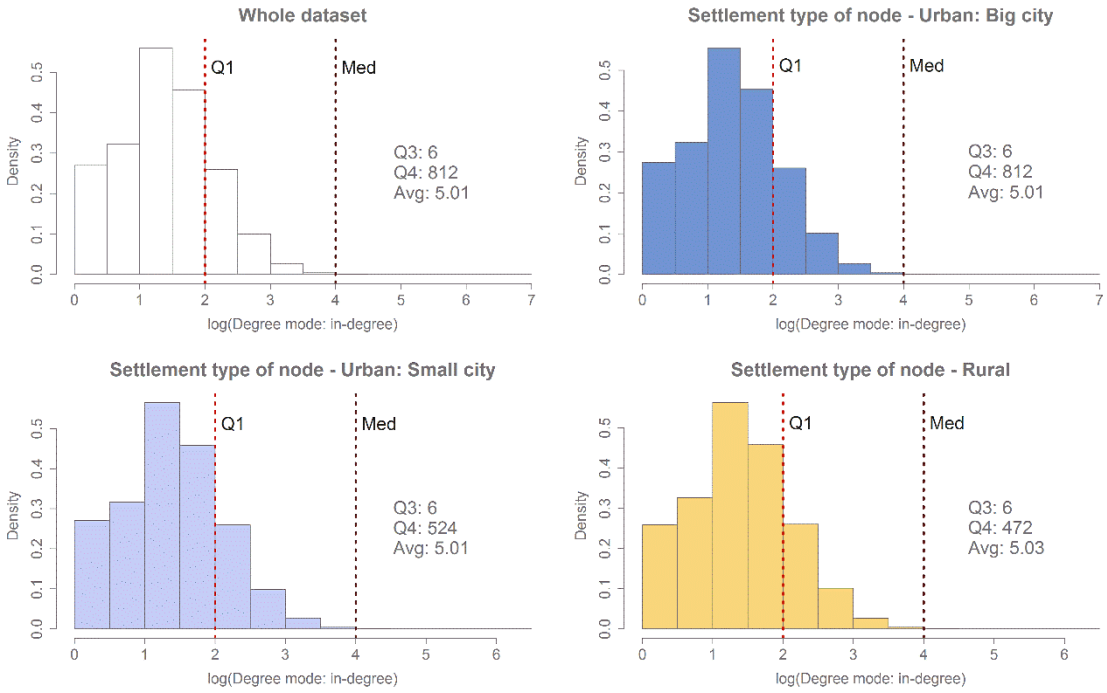


Figure 20. Histograms displaying the node degree by the settlement pairs of links. Mode: In-degree.

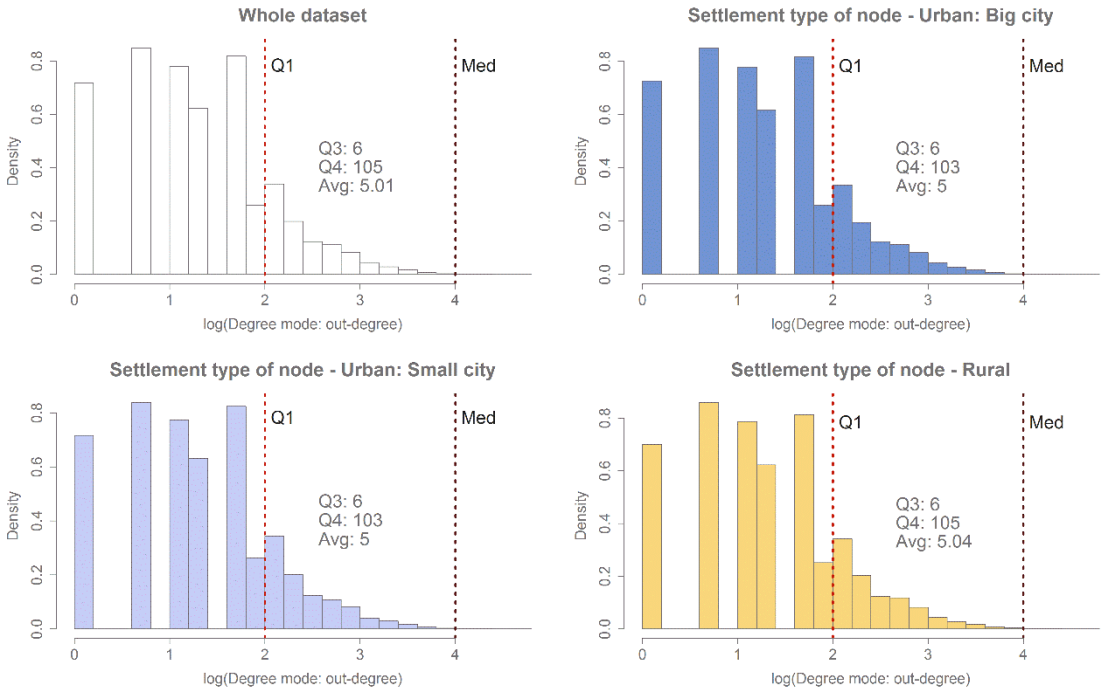


Figure 21. Histograms displaying the node degree by the settlement pairs of links. Mode: Out-degree.

The maximum number of degrees indicates the most central actors of the network; however, it does not demonstrate the distribution range of central nodes. The nodes falling above 99.9% of the values are selected to compare the representativeness by settlement

types. Among the nodes with the highest degrees, 0.1% of the whole dataset of actors, are the nodes with degree values higher than 75, in-degree values taller than 40 and out-degree 44. The most presented nodes are based on big cities by all modes, followed by small cities and rural settlements (see Figure 22). The representativeness of the last two is alike by in-degree and out-degree modes.

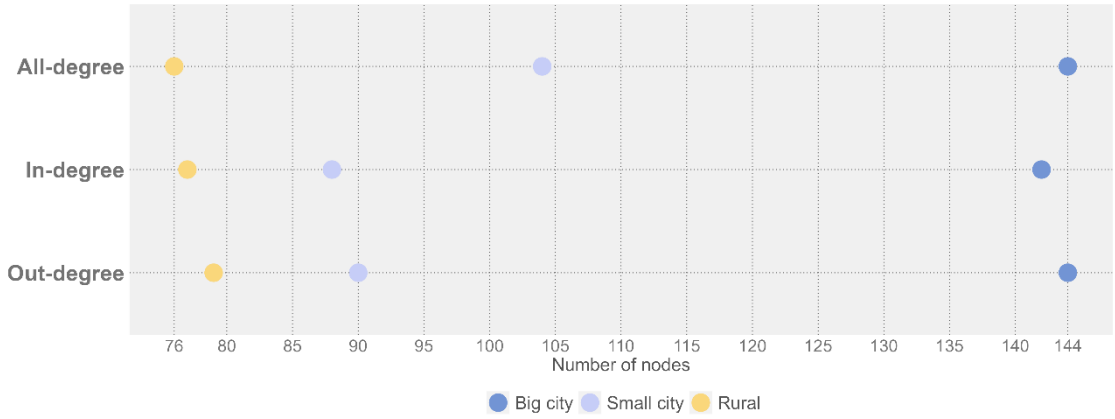


Figure 22. Number of actors by settlement types and degree modes. The actors with higher degree value of 40 are displayed.

Besides the general network centrality analysis, spatial context is also considered to spatially detect the central nodes and settlements. The spatial analysis resulted in revealing exciting patterns. First, the all-degree mode is analysed, and the distribution of nodes is exhibited on the map (see Figure 23). The actors with the highest all-degree values are located in Tallinn and Narva, referring to the assumption that they are the most central actors in the network. Nodes with high all-degree are shown in other big cities. Central nodes by all-degree based on small cities are concentrated around Tallinn, also presented in a few different parts of the country like Kuressaare, Võru, Viljandi etc. (see Figure 23). High node centrality based on rural settlements is shown in a few areas, such as nearby Viljandi and concentration around Tartu.

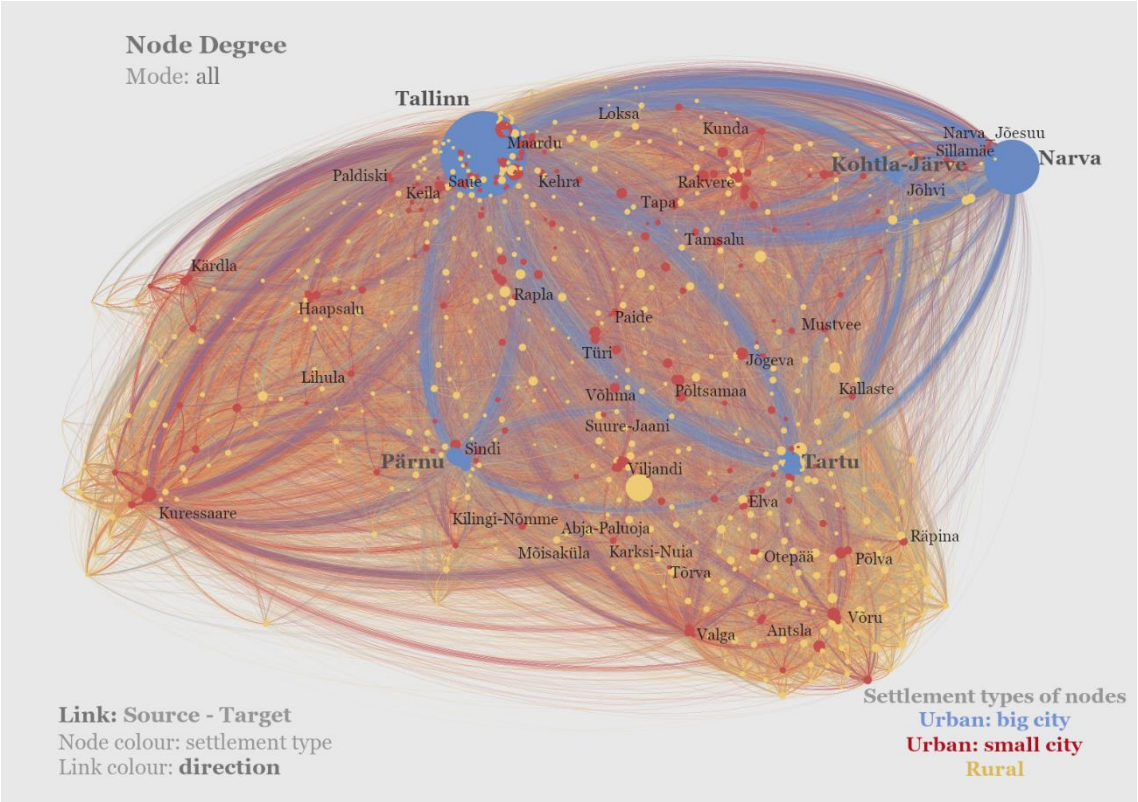


Figure 23. Spatial distribution of node degree. The map shows the degree of nodes based on antenna by the settlement types. The map is created in Gephi by using degree calculation function and GeoLayout. Mode: **All-degree**. The links display both, incoming and outgoing calls.

Differences occur by spatial analysis of in-degree and out-degree modes. Nodes with high in-degree centrality are vastly depicted in big cities (see Figure 24-25), with a few exceptions in small cities and rural settlements. In contrast, by out-degree mode, the high node centrality is exhibited not only for big cities but also for small cities and rural settlements (see Figures 26-27). Furthermore, node centrality based on outgoing calls is higher for nodes in some small cities compared to some nodes in big cities. For instance, Kuressaare and Jõgeva. Another important difference is that node centrality for actors in rural settlements is significantly higher based on outgoing calling activity (see Figures 26-27) than centrality measured based on incoming calls (see Figures 24-25).

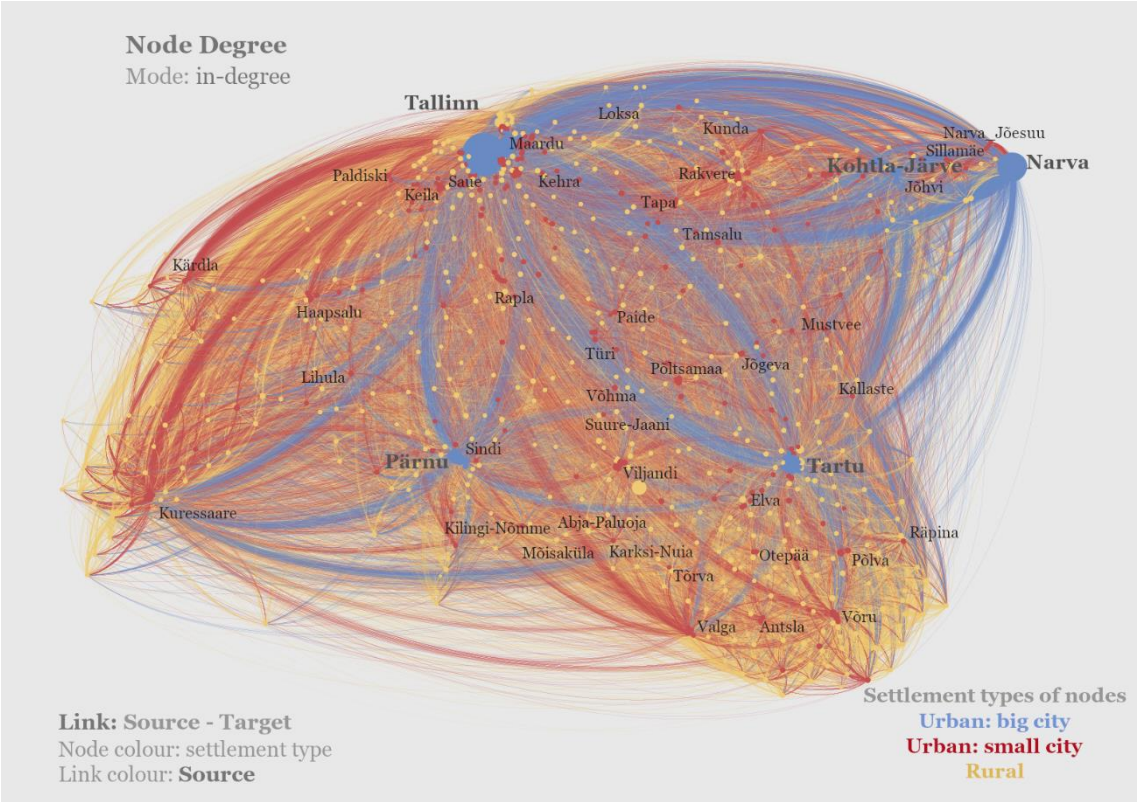


Figure 24. Spatial distribution of node degree. The map shows the degree of nodes based on antenna by the settlement types. The map is created in Gephi by using degree calculation function and GeoLayout. Mode: **In-degree**. The links display outgoing calls.

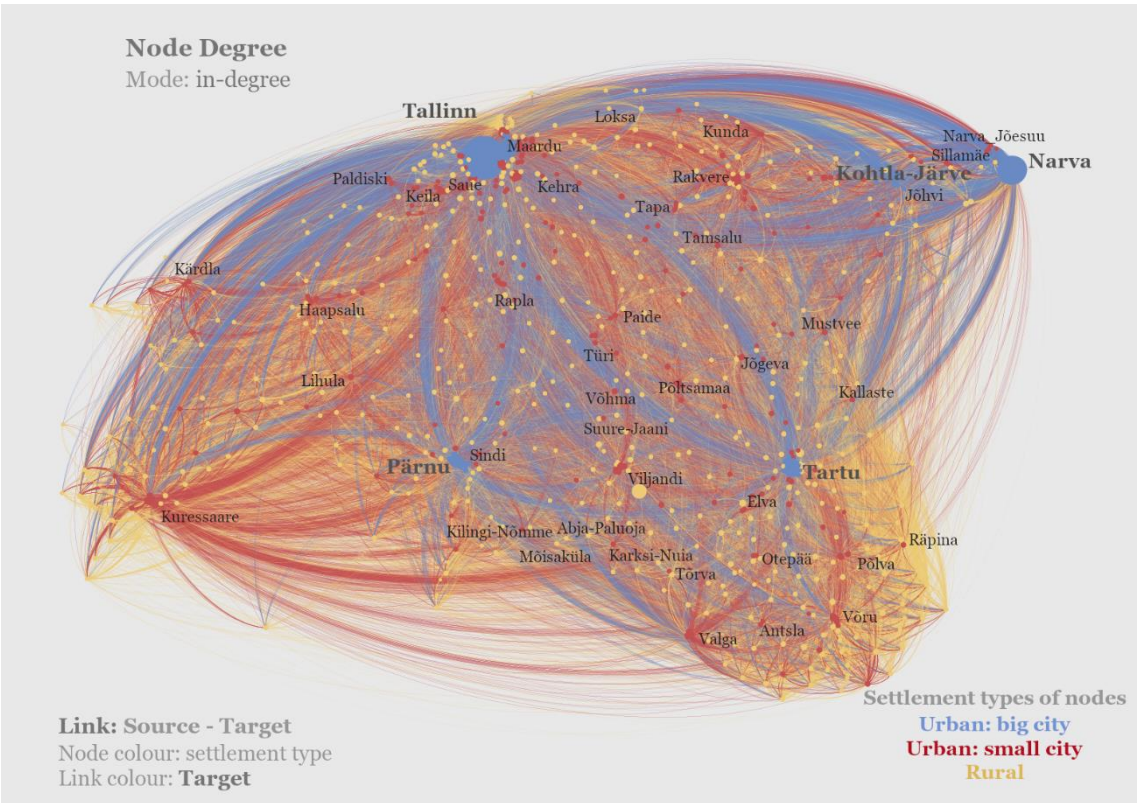


Figure 25. Spatial distribution of node degree. The map shows the degree of nodes based on antenna by the settlement types. The map is created in Gephi by using degree calculation function and GeoLayout. Mode: **In-degree**. The links display incoming calls.

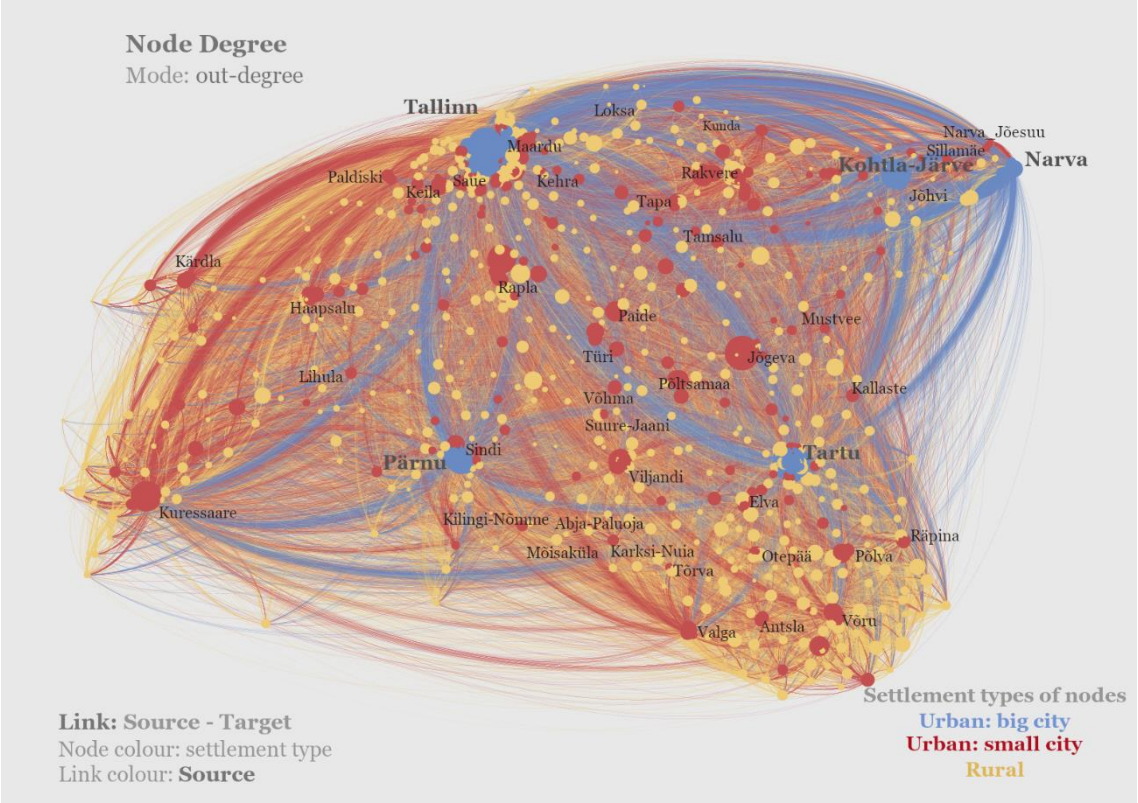


Figure 26. Spatial distribution of node degree. The map shows the degree of nodes based on antenna by the settlement types. The map is created in Gephi by using degree calculation function and GeoLayout. Mode: **Out-degree**. The links display outgoing calls.

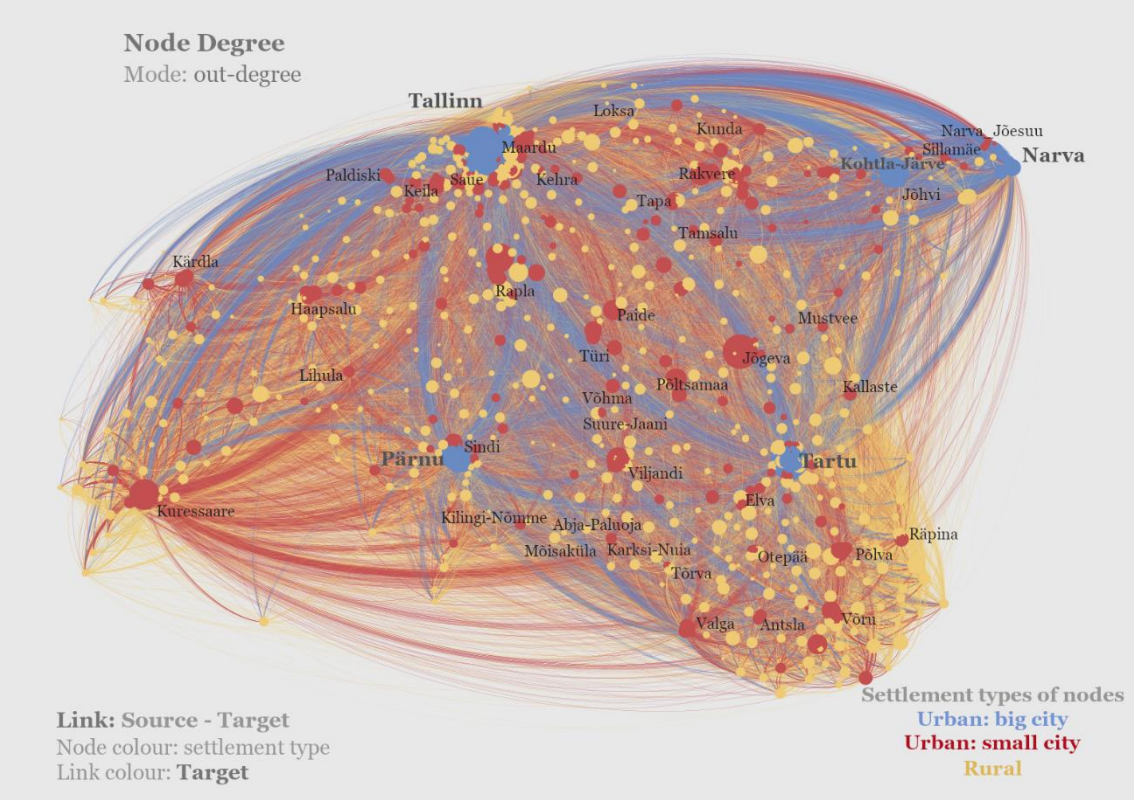


Figure 27. Spatial distribution of node degree. The map shows the degree of nodes based on antenna by the settlement types. The map is created in Gephi by using degree calculation function and GeoLayout. Mode: **Out-degree**. The links display incoming calls.

4.3. Weak and strong ties

The strength of the ties is analysed with the combination of two criteria, weight, which refers to the frequency of calls between call pairs, and edge betweenness centrality described in the methodology part. The derived values of edge betweenness calculations are presented in Table 6. The minimum value of the edge betweenness estimation is 0. The links with a value of 0 indicate the circumstances when no other paths (links) travel through the given edges. The share of links with zero edge betweenness is approximately 20%, while the lower quartile falls at 1. The median value stands at 30, meaning that 30 shortest paths are going through the link with this value. The upper quartile number is 82% higher than the median. The tall maximum value also results in a higher average number, indicating an average of 180 shortest paths travelling through a link of calling pair.

Table 6. Edge betweenness values.

Minimum		0.0
	20%	0.0
Percentiles	25% (Q1)	1.0
	50% (Median)	30
	75% (Q3)	163
	90%	489
	95%	856
	99%	2074
	99.50%	2728
	99.90%	4476
	99.95%	5261
	99.99%	7533
Maximum		18342
Average		180

The weight calculation discussed earlier has revealed that, on average, the call frequency is 3.8 and the upper quartile lies at 4 (see Table ?). The links with a weight above four are considered strong ties with a range of edge weights between 4 and 532. The selected ties share 19.7% of all the links. The strength of the ties is defined due to the frequency of interaction. However, the edge betweenness values are observed for identifying the important links in terms of the frequency of shortest paths through them. As disclosed, all the links have been assigned a betweenness value of 0 as a result of calculation.

The spatial distribution of the strong ties shows that links between the source and target are created at close and far-flung distances (see Figure 28). Nearly 30% of the share of strong ties comes from links created between and within big cities. The map displays the established links by the actors between and within the pairs of settlements, displaying the spatial extent of strong social links on the one hand, though, on the other hand, representing the links that are not important for the network itself.

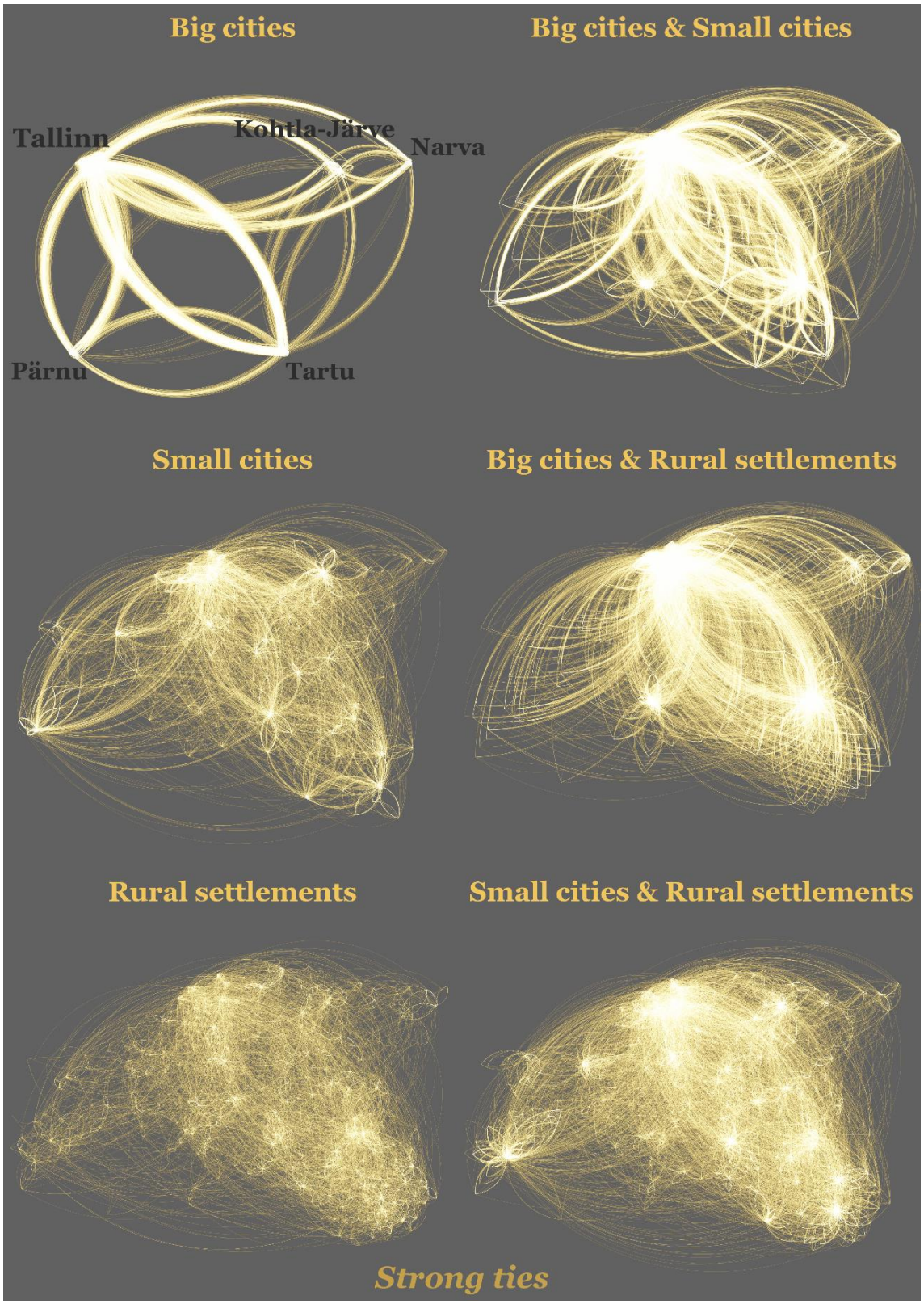


Figure 28. Spatial distribution of strong ties by the settlement types of links.

The ties with a weight (call frequency) of 1 are defined as weak ties. The share of weak ties in the whole dataset is approximately 49%. The edge betweenness values for weak ties vary from 1 to 18,324, with nearly 55% falling between 100 and 1000, and a small

share of the links falls at one and also, above five thousand (see Figure 29). Notably, 0 values of betweenness are not presented for weak ties.

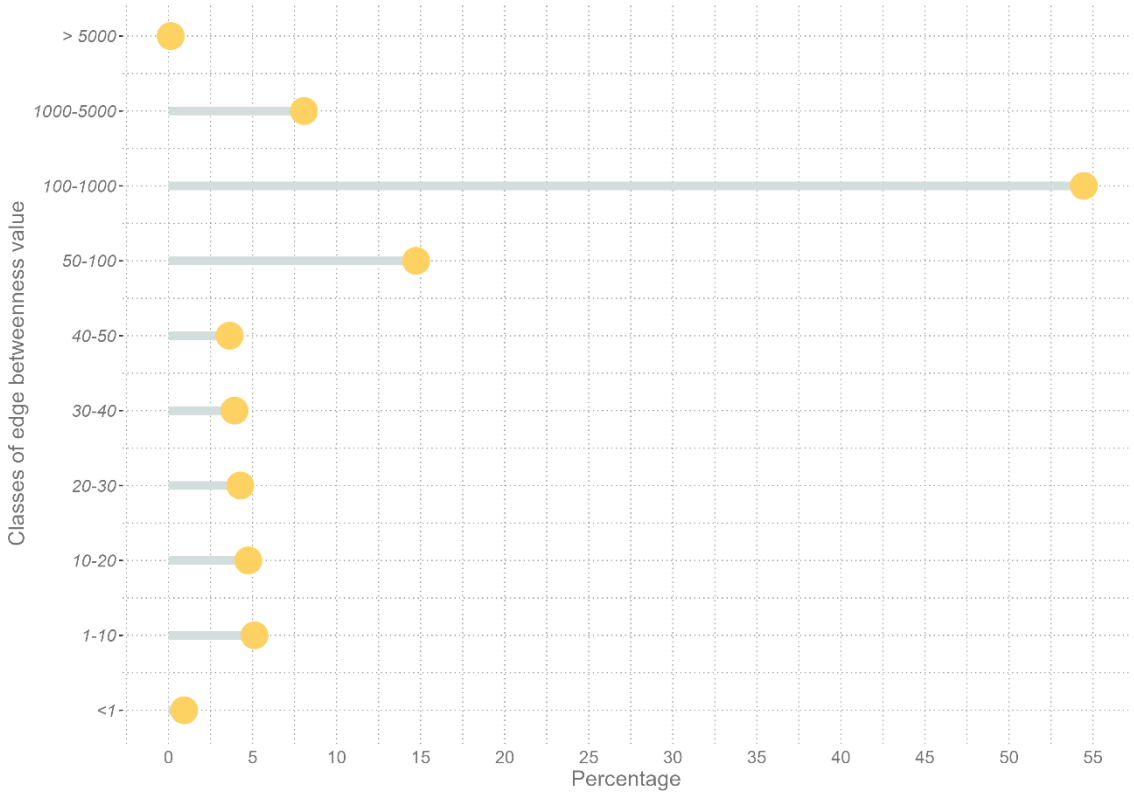


Figure 29. Share (%) of edge betweenness classes.

Displaying the share of weak ties by pairs of settlements shows similar trends by all the pairs - edge betweenness values are primarily distributed within the same range class, with the highest share for links based on big cities and the least for rural settlement pairs (see Figure 30).

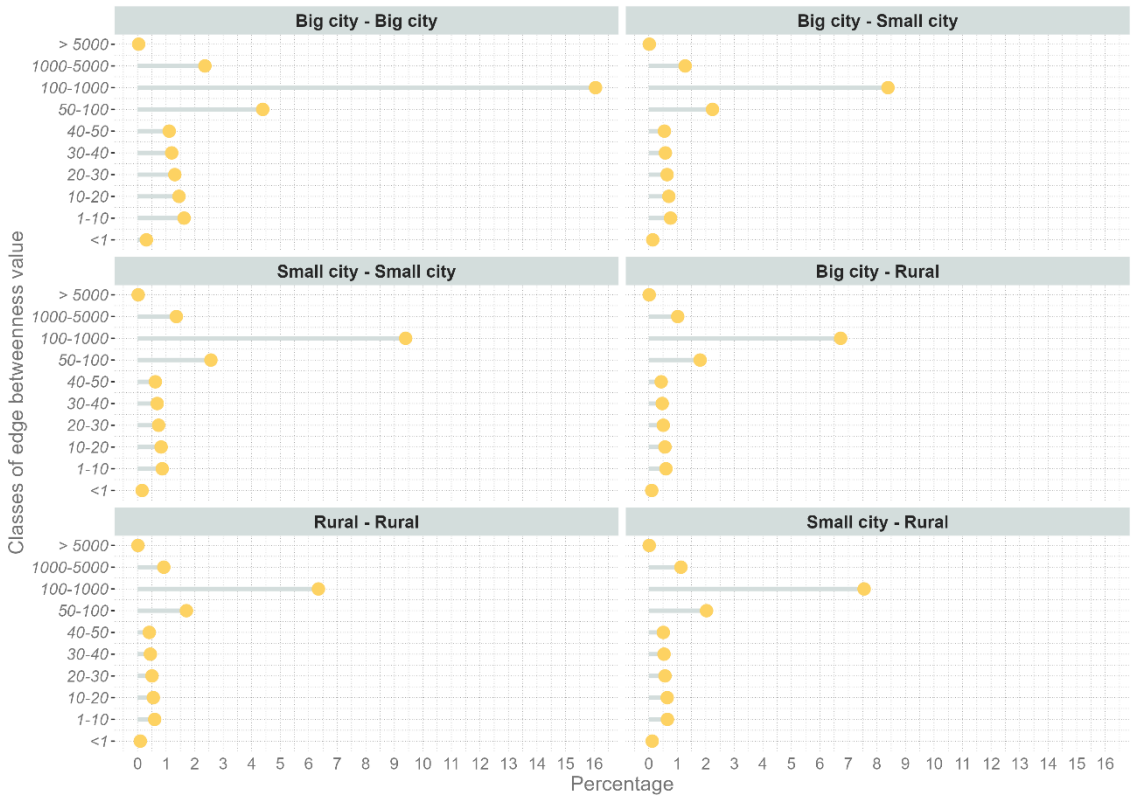


Figure 30. Share (%) of edge betweenness classes (%).

The ties with the highest edge betweenness values are filtered and applied to analysis to estimate their importance in the network. The links with edge betweenness values over 10,000 are selected, representing less than 0.01% of the whole dataset. The selected dataset contains 43 links representing weak ties with a weight of 1 and with no reciprocal relationships. Among the selected links, 27.9% come to the pair of actors based in big cities, following links of actors created between big and small cities, 20.9%, big cities and rural settlements, 18.6%, links between or within small cities - 16.3%, pairs of small cities and rural settlements - 11.6%. A minor share is for the ties of actors created between or within rural settlements, 4.7%.

The selected weak ties (43 links) might be considered from the perspective of bridging different groups and networks; as a result of bridging various connected compounds, the size of the network increases. Further connections are identified to demonstrate the impact of weak ties, considered bridgers, on the network. First, the ties linked to the given 43 links are detected, resulting in a network size of 1,611 calling partners. Increasing the depth of connections and extending the connected network (1,611 links) with other connected ties increases the network size to 8,474 pair links (see Figure 31). With a depth of 3, the network size increases significantly, making up the network of 105,009 calling partners. The connected network includes links with all types of settlement pairs. Thus, the significant impact of weak ties is shown.

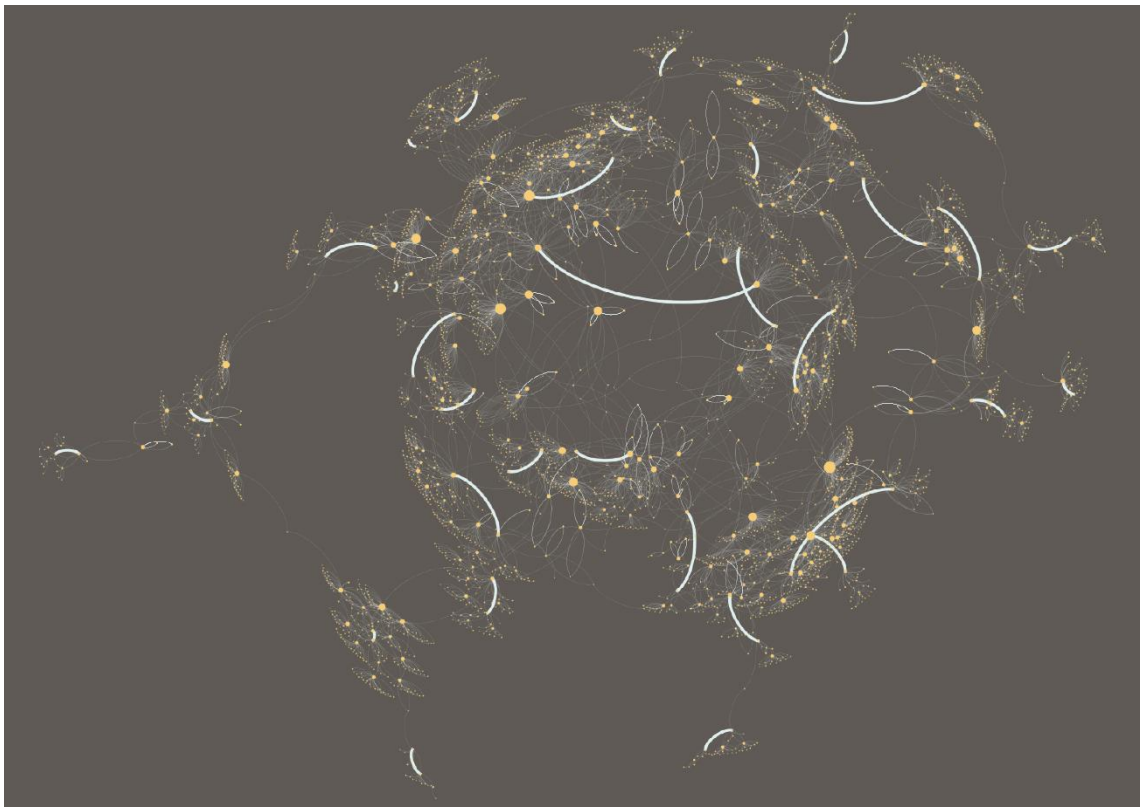


Figure 31. Social network of weak ties (43 selected links). The number of intermediaries in the network is at most 2. The network is created in Gephi by using Yifan Hu Multilevel layout.

The importance of weak ties lies in bridging different groups, especially when groups or networks are homogenous by nature. Figures 32-33 depict the example of linking different groups. The groups consist of actors mainly based in similar types of settlements. Also, the groups differ in size. Connecting such networks is especially important to ensure the spreading of different information.

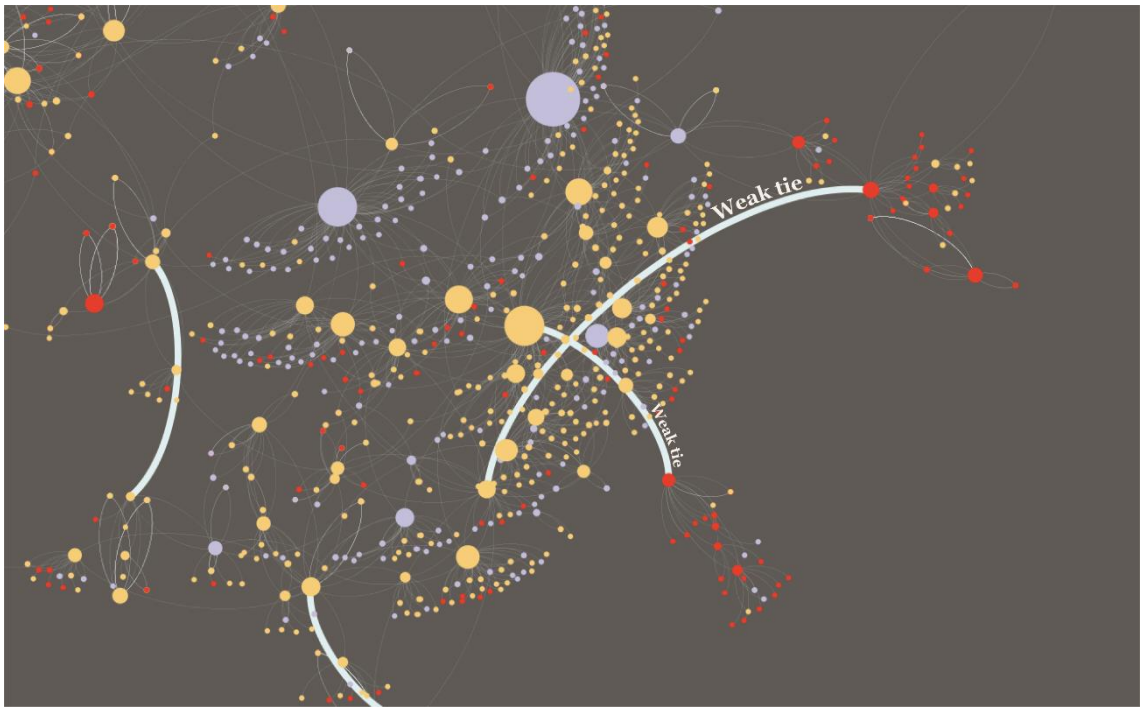


Figure 32. Social network of weak ties (43 selected links). The number of intermediaries in the network is at most 2. The network is created in Gephi by using Yifan Hu Multilevel layout. Zoomed at weak ties. Example 1.

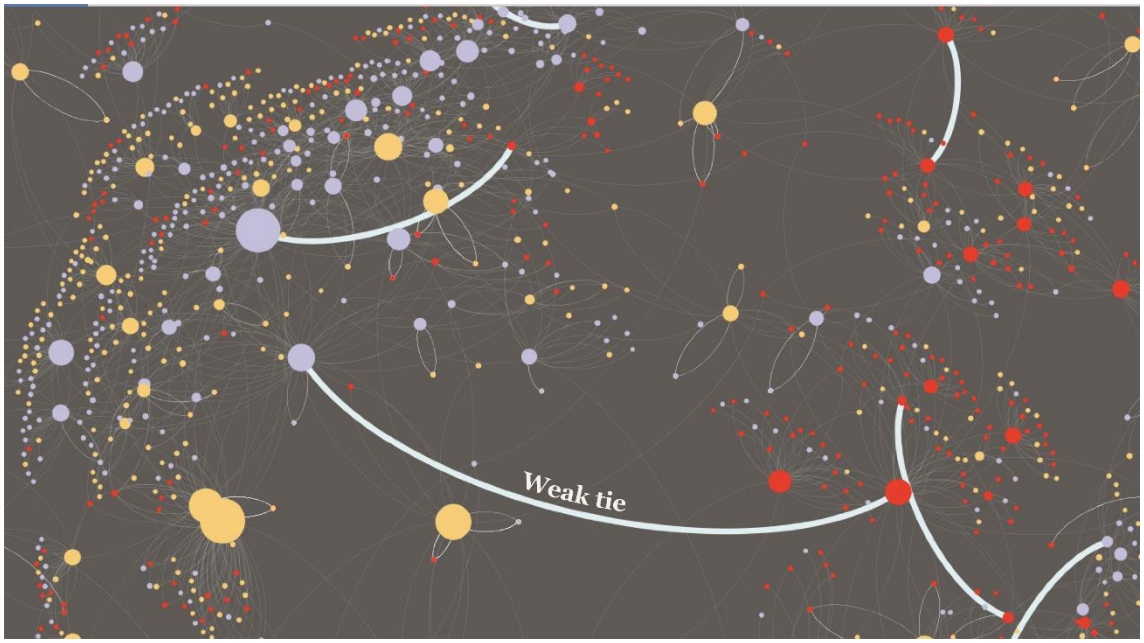


Figure 33. Social network of weak ties (43 selected links). The number of intermediaries in the network is at most 2. The network is created in Gephi by using Yifan Hu Multilevel layout. Zoomed at weak ties. Example 2.

To demonstrate weak ties' power at the individual level, the link with the highest edge betweenness value is selected, and its connected links are identified. The selected weak tie is connected to 34 other links, which themselves are connected to other links, making

up a total of 144 connected ties (see Figure 34). A further step of connected links (depth 3) results in a network of 1,725 connected social ties through, at most, three intermediaries (see Figure 35).

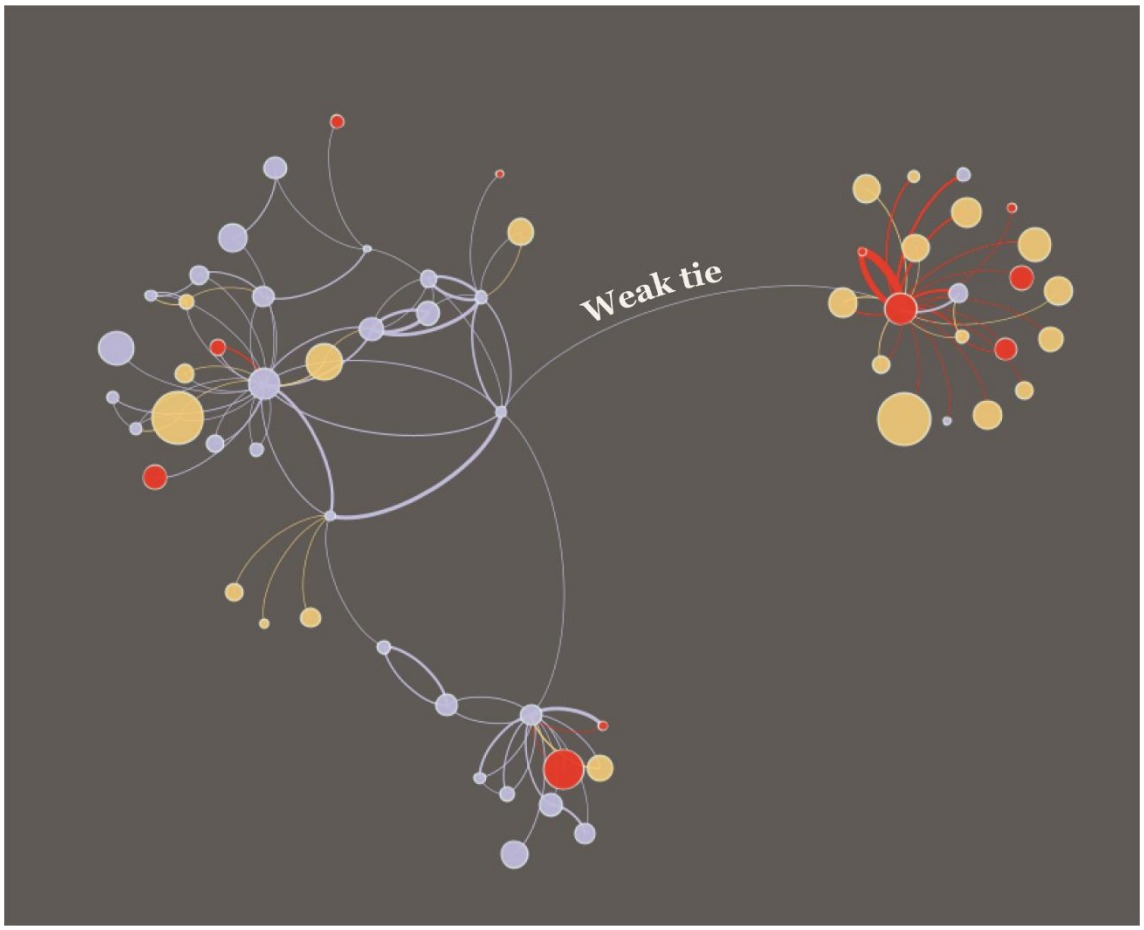


Figure 34. Social network of weak ties with the highest number of the edge betweenness. The number of intermediaries in the network is at most 2. The network is created in Gephi by using Yifan Hu Multilevel layout.

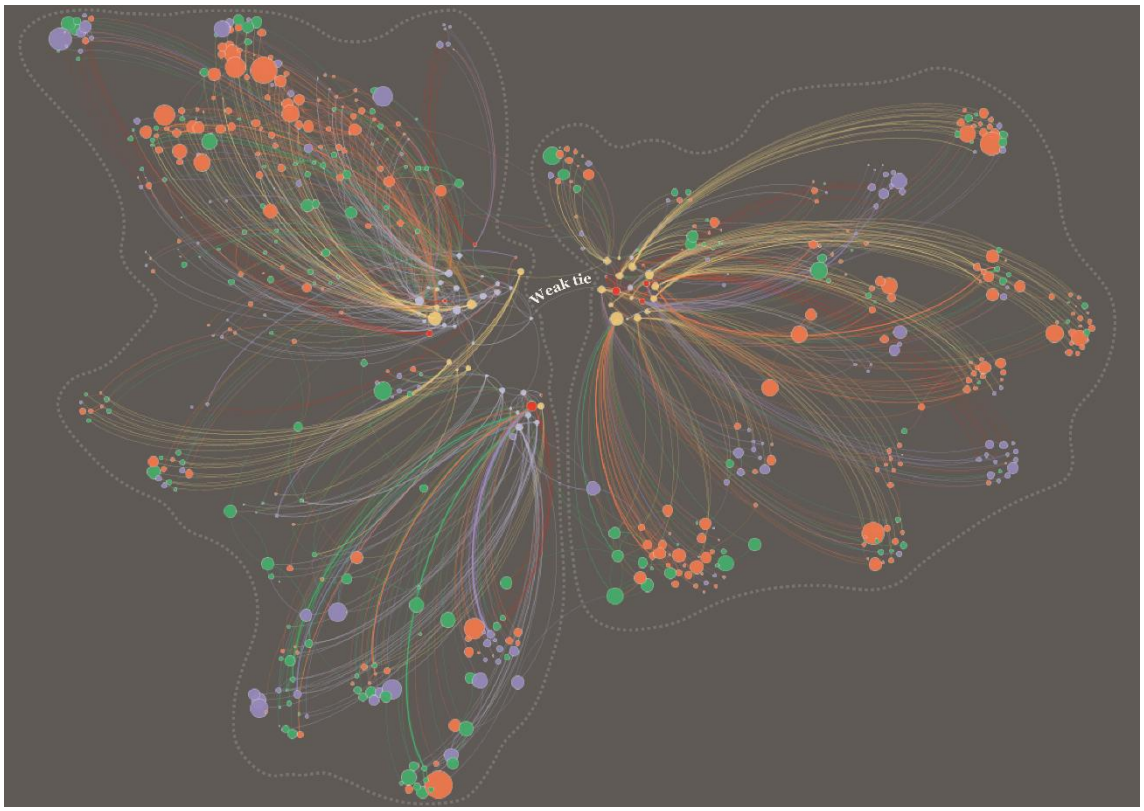


Figure 35. Social network of weak ties with the highest number of the edge betweenness. The number of intermediaries in the network is at most 3. The network is created in Gephi by using Yifan Hu Multilevel layout.

The spatial distribution of weak ties (43 links) and their network connected through, at most, three intermediaries reveal that weak ties are mainly created at far-flung distances (see Figure 36). The weak ties are mostly presented between the northern and southern and northern and western parts of the country.

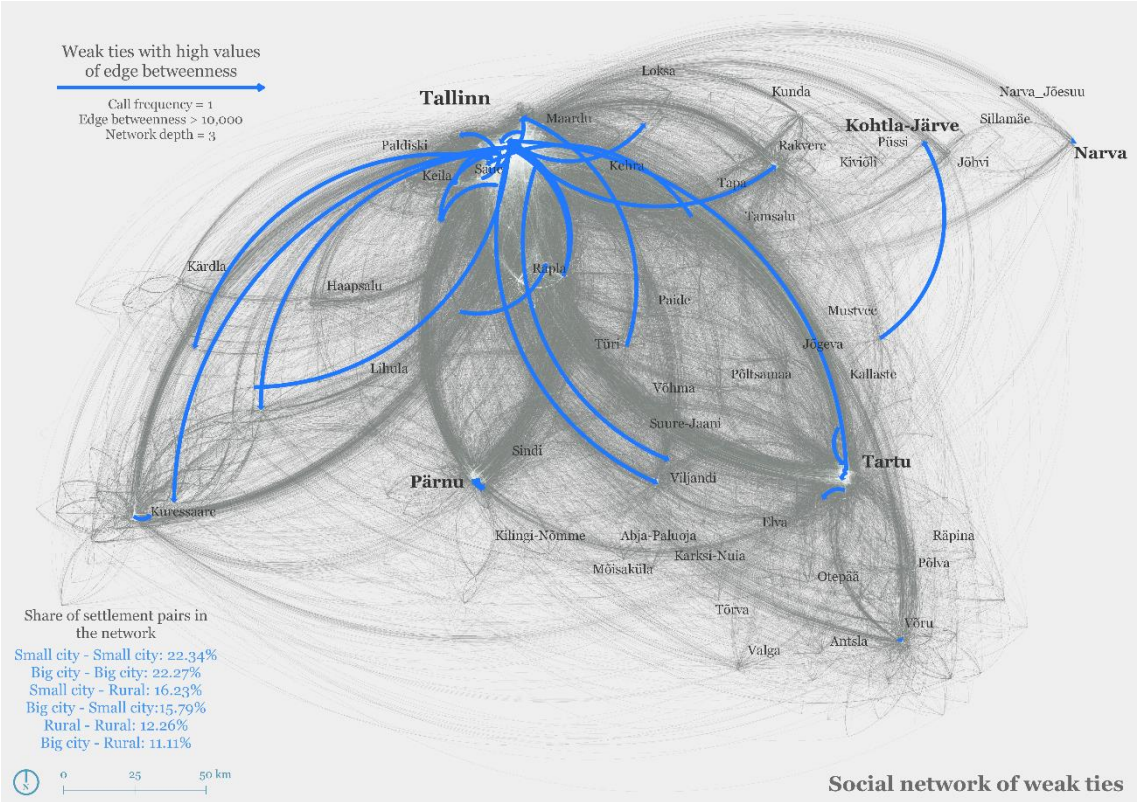


Figure 36. The spatial distribution of the social network of weak ties (43 selected ties). The number of intermediaries in the network is at most 3.

The links with edge betweenness of zero refer to the circumstances when no paths travel through the giving edge, and at least one pair member has only one connection source. Having only one source for interaction and gaining information makes the actor vulnerable. Nearly 22% of the social network of call pairs represent such ties (see Figure 37).

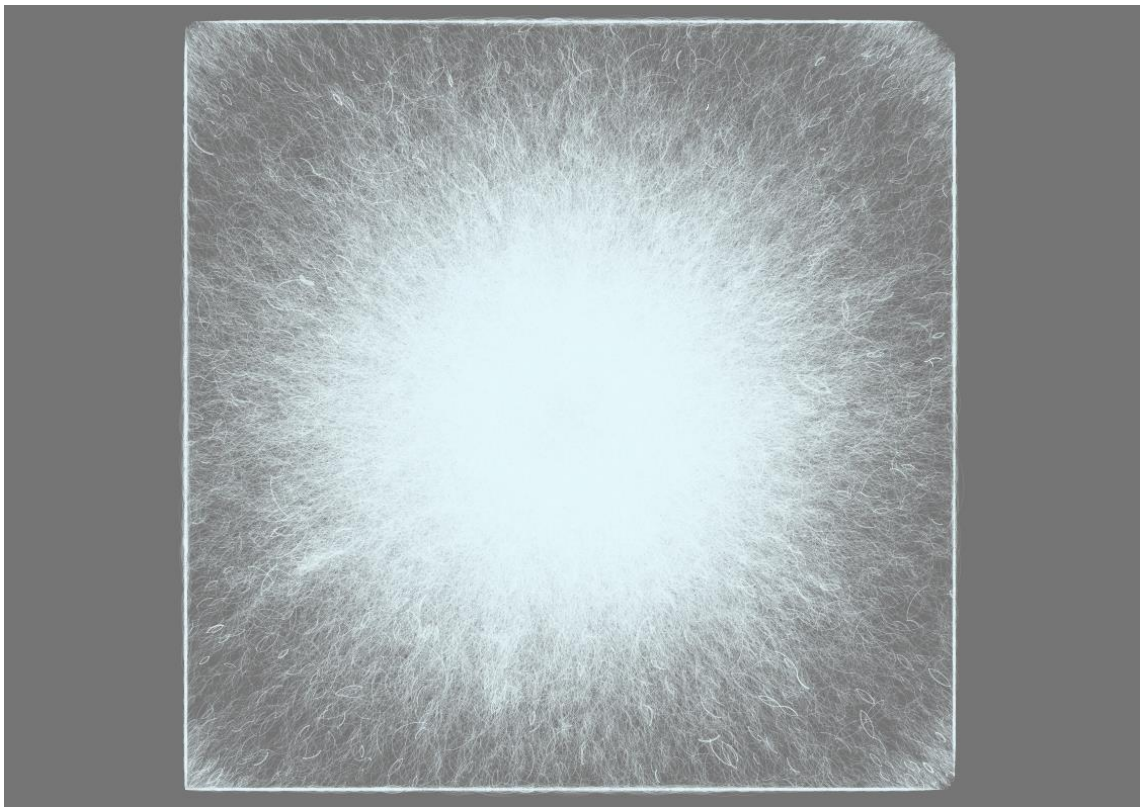


Figure 37. The ties with the edge betweenness of zero.

5. Discussion and conclusions

This study illustrates the combination of network and social sciences and engages the recent trend of implementing digital data into social studies (Lazer et al., 2021). The passive mobile phone dataset is utilised by applying technical tools and analysis methods, and social phenomena are observed by deriving the information from the mobile call dataset. The study reveals, on the one hand, the characteristics of network science due to its data-driven nature (Barabási & Pósfai, 2016; Hansen et al., 2020) and applied instruments to analyse network structure and dyadic relationships (Brandes et al., 2013); on the other hand, the stress is on social facets. The benefits of incorporating both for gaining knowledge about society are demonstrated throughout the study.

The exploratory part of the analysis has revealed that the social ties are mainly created by the actors based on the same settlement types. The settlement types establish the hierarchical relationship. The remarkable pattern in the hierarchy of settlement types lies in the differences in relationships between different types. Representation of established links is more significant between big and small cities, followed by pair of big cities and rural settlements and then by small cities and rural settlements, compared to the pairs of the same types. The reciprocal relationships between settlements are also exhibited. The fact that more social links are created within or between urban settlements, especially big cities, is related to the population distribution and density which is higher in big cities. Other studies also support this assumption, illustrating the impact of population density on dyadic communication, though with a focus on communications within the cities and their peripheries and revealing that dense urban areas support the complementarity of two modes of interaction, face-to-face communications and calling activities (Büchel & Ehrlich, 2020). This study reveals the prospect for further studies on the impact of population density on far-flung dyads and identifying the spatial mobility patterns of dyads in terms of face-to-face interactions.

The spatial distribution of established social links revealed patterns such as the grouping and concentration of calling partners around particular settlements, the presence of far-flung ties and far-flung ties connecting the settlements and the groups of ties around them. The far-flung ties might be intriguing to study from the perspective of internal migration, revealing the spatial patterns of kinships with the stress on social investment. Social investment is required to prevent the decay of relationships and keep active ties (Dindia & Canary, 1993; Burt, 2000; cit. Roberts & Dunbar, 2011) at far-flung distances.

Temporal analysis of the social network of call pairs shows different patterns of calling activities based on weekdays by the settlement type pairs. The social interactions within the same settlements express different characteristics during the weekdays compared to interactions between settlements. The differences occur depending on the pairs of settlement types. Considering the distinct social life rhythm in urban areas and rural settlements, it might be interesting to study the temporal patterns of the social network. Furthermore, to analyse the impact of diurnal rhythm on establishing or keeping the social ties between the settlement types with different social life rhythms and lifestyles. This could also be addressed with the above-mentioned social investment problem.

Network centrality analysis, considering both incoming and outgoing directions of calls, revealed that the most central actors in the network are located in big cities. However, different patterns are demonstrated by estimations based on each mode. Actors' importance and centrality for nodes based in big cities are higher with incoming calls and lower with outgoing calls. In contrast, for small cities and rural settlements, outgoing calls define the importance and centrality of actors. These patterns refer to the circumstances that actors based in the big cities express less initiation to interactions. In contrast, the actors based in small cities and rural settlements reveal more initiatives in communications.

Network centrality might be further studied in the context of segregation. The spatial distribution of node degree shows that some important actors are located in Narva, with primarily a Russian-speaking community and the highest representativeness of a Russian-speaking community in Estonia, followed by Tallinn (Silm et al., 2021). Further studies could be conducted on the role of actors' leadership between Narva and other settlements populated by the Russian-speaking community to analyse the spatial patterns of personal influence and its impact on information dissemination in social circles. Moreover, the actors' influence and the impacts of information dissemination, based on ethnolinguistic characteristics, might be applied in the applications such as studying the spatial patterns of electoral behaviour, the spread of conspiracy theories, demographic patterns of the actor leadership, etc.

Analysing the strength of social ties exhibits different patterns. Detected strong ties represent a one/fifth of the whole network. Though, through the given links, no shortest paths of other links travel, making these social ties closed since one pair member has only one interaction source. On the one hand, these links indicate strong connections between two actors, though, on the other hand, they might be discussed from the perspective of vulnerability. The detected strong ties are assumed to represent kinship and close friendship ties since they contribute more to the relationship, according to Granovetter (1973). The spatial distribution patterns of strong ties also sustain the hierarchical characteristics discussed before. Among weak ties, the edges with high importance are detected. It is revealed that some weak ties have significant importance as they connect different groups and networks. The significant importance of weak ties is seen in connecting small groups of nodes with the same settlement types to a bigger group with different settlement types.

The study has limitations that need to be noted. The utilised mobile phone data represents only one mobile network operator. Thus, analysing the dataset containing one network provider excludes the ties created by the users of different network providers. This circumstance might affect the biased outcomes. The absence of attributes of age, gender and ethnolinguistic composition makes it hard to relate the patterns of social networks to other related studies.

The high number of mobile cellular phone subscriptions (Central Intelligence Agency, 2022), meaning that a person might have more than one mobile phone device, can result in duplicated users in the dataset and affect biased results and assumptions.

The methodology to detect the home locations is partly based on considering the diurnal rhythm of hours spent at home and work. Though, the tendency of teleworking, which occurred before the pandemic period, might affect the detection of home anchors. The trend of remote working, part-time working, freelancing etc., makes it hard to define the location of home based on the diurnal rhythms (Tench et al., 2002).

Technological development has changed the ways of communication and interaction. Also, young people have adapted to typing-based communications (Wellman, 2001). The usage of mobile calls for interactions might have age-based patterns. Friendship ties depend not only on mobile calls but also on social media. The analysed social network of call partners might primarily refer to kinships.

The utilised data represents the period before the pandemic. Thus, the assumptions and estimations might not correspond to the current trends impacted by the pandemic. The issue of social distancing might impact social networks and could be the focus of further research.

6. Summary

The study's objectives were to explore the social network of call pairs, identify the essential patterns, and answer the central questions of network centrality and the strength of the ties. The study aimed to bring new knowledge to the field by addressing the spatial context with the social network. The mobile phone dataset was utilised, which required the implementation of network science and spatial analysis tools. The exploratory analysis resulted in various aspects and demonstrated the further prospects of the research. Analysing network centrality and the strength of the ties resulted in more specific outcomes. The central and the most important actors are based in big cities, followed by small cities. The spatial patterns reveal that the spread of influence and information dissemination has a large extent, including the fur-flung relationships. Also, the influence and the spread of information reaches not only the same settlement types but different types. It is noticeable that the incoming call from rural settlements highly defines the centrality of actors based in big or small cities. Actors based in rural settlements reveal that outgoing calling activity is higher than incoming calling activity, indicating that they make an effort to maintain ties with actors in urban areas.

Analysing the strength of the ties revealed that nearly 20% of links are strong, presumably representing kinships. The spatial distribution of strong ties corresponds to the hierarchical patterns of the relationships based on settlement types. The strong ties at the same time represent the vulnerable links due to the limited source of information. Strong ties correspond to the pairs, where at least one actor has only one connection. The information dissemination within such links is not prone to be diverse, though the strength lies in maintaining the ties.

On the other hand, nearly 22% of the network consists of weak links. Among weak links, the importance was measured by the number of shortest paths travelled through. The select weak ties with the highest importance values (edge betweenness) demonstrated the role of weak ties for information dissemination in the network. The selected 43 weak ties created a network of 105,009 social links connected through, at most, three intermediaries. Thus, weak ties are seen as the source of spreading diverse or different information within this network. While one weak tie (with the highest importance values) connected 1,725 links through three intermediaries. Information dissemination is essential between different settlement types. The spatial distribution of the social network connected by the 43 weak ties through three intermediaries reveals more concentration between the northern and southern and northern and north-western parts. In contrast, the northeastern parts, including Narva and its surrounding, is less connected.

This study has limitations lying in data, methodology and implementation. However, it has demonstrated the value of network analysis in studying social networks and has shown the prospects of further studies.

Sotsiaalvõrgustikud ja informatsiooni levitamise ruumilised mustrid passiivsete mobiilpositsioneerimise andmete põhjal

Lika Zhvania

Kokkuvõte

Sotsiaalvõrgustikud on inimestest, sugulus- ja sõprussidemetest või muudest sotsiaalsetest suhetest koosnevad keerulised süsteemid. Sotsiaalvõrgustikes levitatakse, vahetatakse ja vahendatakse informatsiooni, teadmisi, ressursse, käitumisjooni ja mõjuvõimu. Võrgustiku struktuur defineerib selle, millist tüüpi ühendused on esindatud omavahel tihedalt läbi käivate inimestega ja alamvõrgustikega ja millised grupid on informatsiooni leviku seisukohast haavatavad. Uurimistöö lisab uusi teadmisi sotsiaalvõrgustike uurimisel eelkõige esile tuues sotsiaalvõrgustike ruumilist külge, kasutades selleks mobiilpositsioneerimise andmeid.

Uurimistöö eesmärkideks on analüüsida telefonikasutajate suhtlusvõrgustikke lähtudes telefonikasutajate peamise asustusüksuse tüübist ja tuvastada võrgustiku ruumilised mustrid Eestis. Uurimistöös on kasutatud andmepõhist lähenemist, et kirjeldada võrgustiku struktuuri kogu riigis asustusüksuse täpsusega.

Töös on kaks peamist uurimisküsimust:

1) Millised on võrgustiku kesksuse ruumilised mustrid ja tunnused? Milline on tähtsate asustusüksuste muster asustustüüpide lõikes?

2) Milline on tugevate sotsiaalsete sidemete ruumiline muster? Millised on tugevad ja nõrgad sidemed ning vajalikud ühenduslülid?

Uurimistöö peamisteks teoreetilisteks lähtekohtadeks on võrgustiku kesksus ja nõrkade sidemete tugevus.

Kesksus võrgustikus on üks peamiseid sotsiaalvõrgustiku struktuuri kirjeldavaid tunnuseid ja viitab võrgustiku liikme olulisusele võrgustiku sees. Sotsiaalvõrgustikes nähakse kesksust kui mõjuvõimule viitavat tegurit, mida saab siduda sotsiaalsete rollidega nagu juhtimine (Leavitt, 1951; Berkowitz, 1956; Shaw, 1964; cit. Bonacich, 1987).

Sotsiaalvõrgustike sidemete tugevus on kombinatsioon tunnustest nagu suhtlemise tihedus, emotsionaalne lähedus, intiimsus ja vastastikune aktiivsus (Granovetter, 1973). On keeruline tõmmata piire selle vahel, kas võrgustiku side on tugev, nõrk või puudub. Kui kahe inimese vaheline suhtlusaktiivsus on madal, mitte-lähedane või ühepoolne, peetakse enamasti nendevahelist sotsiaalset sidet nõrgaks. Omavahel suguluses olevaid inimesi ja lähedasi sõpru loetakse tugevateks sidemeteks. Sellegipoolest on ka nõrkadel sidemetel oma tugevus, mis väljendub informatsiooni edastamises ja mõjus. Kui gruppide sees on enam-vähem sama palju tugevaid sidemeid ja levitatakse sarnast informatsiooni, siis tekib rohkem uudseid ideid nendes gruppides, millel on rohkem nõrku sidemeid teiste gruppidega.

Uurimistöös on kasutatud passiivseid mobiilpositsioneerimise andmeid, et analüüsida inimeste kõnetoimingutest tekkivadt sotsiaalvõrgustikku. Metoodika loomisel on olulisel kohal suhtlussidemete ühendamine ruumilise kontekstiga, mis võimaldab kirjeldada võrgustiku ruumilisi mustreid.

Uurimistöös on esmalt välja arendatud metoodika andmete puhastamiseks ja töötlemiseks, et mobiilpositsioneerimise andmetel ruumilist võrgustikuanalüüsi läbi viia. Teiseks on kasutatud mitmeid võrgustiku analüüsi meetodeid ja võimalusi, et kirjeldada ning visualiseerida töö tulemusi.

Töö tulemusena selgub, et sotsiaalsed sidemed on enimlevinud samas asustusüksuse tüübis asuvate võrgustiku liikmete vahel. Kui jätta kõrvale asustustüüpide sisesed sidemed, siis on kõike sagedasemad sotsiaalsed sidemeid suurte linnade ja väikeste linnade vahel, seejärel suurte linnade ja maapiirkondade vahel ning kõige vähem on sidemeid väikeste linnade ja maapiirkondade inimeste vahel. Kõnetoimingute ruumiline analüüs võimaldab kirjeldada võrgustikusisest suhete kontsentreerumist, grupeerumismustreid ja vahemaaüleseid sotsiaalseid sidemeid asustusüksuste vahel. Kõnetoimingute ajaline muster on asustustüüpide vahel nädalapäeviti erinev.

Lähtudes sissetulevatest ja väljaminevatest kõnedest asuvad võrgustiku kõige kesksemad lülid suurtes linnades. Samas erinevad ruumilised mustrid erinevates kesksuse tüüpides. Suurtes linnades asuvate võrgustikuliikmete kesksus on suurem sissetulevate kõnede põhjal. Seevastu väiksemates linnades ja maapiirkondades on väljaminevaid kõnesid rohkem.

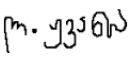
Sotsiaalse võrgustiku sidemete tugevused näitavad mustreid teise nurga alt. Ühe viiendiku kõigist võrgustiku sidemetest võib defineerida kui tugevad sidemed. Nende hulgas on võrgustikus välja kujunenud ka suletud grupid, kus võrgustiku liikmetel on tugev omavaheline side, kuid samas puuduvad sidemed teiste gruppidega. See võib viidata liikmetele väikestes gruppides, kelle informatsioonialane sõltuvus teineteisest on suur, mis vääriks edasist tähelepanu teadustöödes nende võimaliku haavatavuse perspektiivis. Seevastu, osad nõrgad sidemed on sotsiaalvõrgustikes olulisteks lülideks, mis ühendavad väiksemad grupid teatud tüüpi asustusüksustes suuremate gruppidega teist tüüpi asustusüksustes.

Acknowledgements

I would like to thank my supervisors, Anto Aasa and Anniki Puura, for introducing me to the topic, assisting with research, providing the data, establishing the agreement for data usage, and ensuring access for working in the workstation. I would like to express my gratitude to the Mobility Lab of the University of Tartu, the holder of the data utilised in the research and manager of the workstation where the whole data analysis took place. I also thank the IT staff who provided technical support during working in the workstation.

I want to thank Gabrielius Baranauskas for his support and encouragement throughout the study.

My thanks,



References

Ahas, R., Aasa, A., Roose, A., Mark, Ü., & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, *29*(3), 469–486. https://doi.org/10.1016/j.tourman.2007.05.014

Ahas, R., Silm, S., Järv, O., Saluveer, E., & Tiru, M. (2010). Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, *17*(1), 3–27. https://doi.org/10.1080/10630731003597306

Barabási, A.-L. (2003). *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*. PLUME. Penguin Group.

Barabási, A.-L., & Pósfai, M. (2016). *Network Science*. Cambridge University Press. http://networksciencebook.com/

Berkowitz, Leonard. (1956). Personality and Group Position. *Sociometry*, *19*(4), 210. https://doi.org/10.2307/2785764

Bonacich, P. (1987). Power and Centrality : A Family of Measures. *American Journal of Sociology*, *92*(5), 1170–1182. https://www.jstor.org/stable/2780000

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, *27*(1), 55–71. https://doi.org/10.1016/j.socnet.2004.11.008

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network Analysis in the Social Sciences. *Science*, *323*(5916), 892–895. https://doi.org/10.1126/science.1165821

Brandes, U. (2001). A faster algorithm for betweenness centrality*. *The Journal of Mathematical Sociology*, *25*(2), 163–177. https://doi.org/10.1080/0022250X.2001.9990249

Brandes, U., Robins, G., McCranie, A., & Wasserman, S. (2013). What is network science? *Network Science*, *1*(1), 1–15. https://doi.org/10.1017/nws.2013.2

Bruggeman, J. (2008). *Social Networks: An Introduction*. Routledge.

Büchel, K., & Ehrlich, M. V. (2020). Cities and the structure of social interactions: Evidence from mobile phone data. *Journal of Urban Economics*, *119*(April). https://doi.org/10.1016/j.jue.2020.103276

Burt, R. S. (2000). Decay functions. *Social Networks*, *22*(1), 1–28. https://doi.org/10.1016/S0378-8733(99)00015-5

Carrington, P. J., Scott, J., & Wasserman, S. (2005). *Models and Methods in Social Network Analysis*. www.cambridge.org/9780521809597

Carron, P. Mac, Kaski, K., & Dunbar, R. (2016). Calling Dunbar ' s numbers. *Social Networks*, *47*, 151–155. https://doi.org/10.1016/j.socnet.2016.06.003

Central Intelligence Agency. (2022). *Estonia*. THE WORLD FACTBOOK. Estonia. Communications. www.cia.gov

Chakraborty, A., Dutta, T., Mondal, S., & Nath, A. (2018). Application of Graph Theory in Social Media. *International Journal of Computer Sciences and Engineering*, *6*(10), 722–729. https://doi.org/10.26438/ijcse/v6i10.722729

Choudhury, P. (Raj), Foroughi, C., & Larson, B. (2021). Work-from-anywhere: The

productivity effects of geographic flexibility. *Strategic Management Journal*, *42*(4), 655–683. https://doi.org/10.1002/smj.3251

Dindia, K., & Canary, D. J. (1993). Definitions and Theoretical Perspectives on Maintaining Relationships. *Journal of Social and Personal Relationships*, *10*(2), 163–173. https://doi.org/10.1177/026540759301000201

Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews*, *6*(5), 178–190. https://doi.org/10.1002/(SICI)1520-6505(1998)6:5<178::AID-EVAN5>3.0.CO;2-8

Easley, D., & Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press. https://doi.org/10.1145/335305.335325

Encyclopedia Britannica. (2022). *Estonia*. www.britannica.com

Freeman, L. C. (1978a). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239. https://doi.org/10.1016/0378-8733(78)90021-7

Freeman, L. C. (1978b). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239. https://doi.org/10.1016/0378-8733(78)90021-7

Gamper, M. (2022). Social Networks and Health Inequalities. In A. Klärner, M. Gamper, S. Keim-Klärner, I. Moor, H. von der Lippe, & N. Vonneilich (Eds.), *Social Networks and Health Inequalities* (Issue June). Springer International Publishing. https://doi.org/10.1007/978-3-030-97722-1

Gephi. (2022). The Open Graph Viz Platform. https://gephi.org/

Goel, R., Sharma, R., & Aasa, A. (2021). Understanding gender segregation through Call Data Records: An Estonian case study. *PLOS ONE*, *16*(3), 1–21. https://doi.org/10.1371/journal.pone.0248212

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360–1380. https://www.jstor.org/stable/2776392

Hale, A. E. (2009). Moreno's Sociometry: Exploring Interpersonal Connection. *JSTOR. The Eastern Group Psychotherapy Society (EGPS)*, *33*(4), 347–358. https://www.jstor.org/stable/41719254

Hansen, D. L., Shneiderman, B., Smith, M. A., & Himelboim, I. (2020). *Analyzing Social Media Networks with NodeXL*. Elsevier. https://doi.org/10.1016/C2018-0-01348-1

Hill, R. A., & Dunbar, R. I. M. (2003). Social network size in humans. *Human Nature*, *14*(1), 53–72. https://link.springer.com/article/10.1007/s12110-003-1016-y

Hoffman, M. (2021). *Methods for Network Analysis*. https://bookdown.org/markhoff/social_network_analysis/

igraph. (2022). The Network Analysis Package. https://igraph.org/

Karinthy, F. (1929). *Chains*. https://short-stories.co/stories/chains-4w0n6QJmNDr

Land Board. (2022). *Spatial data catalogue*. Geoportal. Republic of Estonia. Land Board. https://geoportaal.maaamet.ee/eng/

Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1944). The people's choice: How the voter makes up his mind in a presidential campaign. *NY: Columbia University Press*.

Lazer, D., Hargittai, E., Freelon, D., Gonzalez-Bailon, S., Munger, K., Ognyanova, K., &

Radford, J. (2021). Meaningful measures of human society in the twenty-first century. *Nature*, *595*(7866), 189–196. https://doi.org/10.1038/s41586-021-03660-7

Leavitt, H. J. (1951). Some effects of certain communication patterns on group performance. *The Journal of Abnormal and Social Psychology*, *46*(1), 38–50. https://doi.org/10.1037/h0057189

Licoppe, C., & Smoreda, Z. (2005). Are social networks technologically embedded? How networks are changing today with changes in communication technology. *Social Networks*, *27*(4), 317–335. https://doi.org/10.1016/j.socnet.2004.11.001

Liu, W., Sidhu, A., Beacom, A. M., & Valente, T. W. (2017). Social Network Theory. In *The International Encyclopedia of Media Effects* (pp. 1–12). Wiley. https://doi.org/10.1002/9781118783764.wbieme0092

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*(1), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

Merton, K., R. (1957). *Social Theory and Social Structure*. Free Press, Glencoe, Ill. https://www.worldcat.org/title/4536864

Milgram, S. (1967). The Small-World Problem. *Psychology Today*, *1*(1), 61–67.

Palchykov, V., Kertész, J., Dunbar, R., & Kaski, K. (2013). Close Relationships: A Study of Mobile Communication Records. *Journal of Statistical Physics*, *151*(3–4), 735–744. https://doi.org/10.1007/s10955-013-0705-0

Pool, I. de S., & Kochen, M. (1978). Contacts and influence. *Social Networks*, *1*(1), 5–51. https://doi.org/10.1016/0378-8733(78)90011-4

Puura, A. (2022). *Relationships between personal social networks and spatial mobility with mobile phone data* [University of Tartu]. http://hdl.handle.net/10062/76073

Puura, A., Silm, S., & Ahas, R. (2018). The Relationship between Social Networks and Spatial Mobility: A Mobile-Phone-Based Study in Estonia. *Journal of Urban Technology*, *25*(2), 7–25. https://doi.org/10.1080/10630732.2017.1406253

Puura, A., Silm, S., & Masso, A. (2022). Identifying relationships between personal social networks and spatial mobility: A study using smartphone tracing and related surveys. *Social Networks*, *68*(August 2021), 306–317. https://doi.org/10.1016/j.socnet.2021.08.008

Rainie, L., & Wellman, B. (2014). *Networked: The New Social Operating System*. The MIT Press.

Roberts, S. G. B., & Dunbar, R. I. M. (2011). Communication in social networks: Effects of kinship, network size, and emotional closeness. *Personal Relationships*, *18*(3), 439–452. https://doi.org/10.1111/j.1475-6811.2010.01310.x

Saramäki, J., Leicht, E. A., López, E., Roberts, S. G. B., Reed-Tsochas, F., & Dunbar, R. I. M. (2014). Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences*, *111*(3), 942–947. https://doi.org/10.1073/pnas.1308540110

Shaw, M. E. (1964). Communication Networks. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 1, pp. 111–147). New York: Academic. https://doi.org/10.1016/S0065-2601(08)60050-7

Silm, S., Mooses, V., Puura, A., Masso, A., Tominga, A., & Saluveer, E. (2021). The Relationship between Ethno-Linguistic Composition of Social Networks and Activity Space: A Study Using Mobile Phone Data. *Social Inclusion*, *9*(2). https://doi.org/10.17645/si.v9i2.3839

Statistics Estonia. (2021). *Census 2021. Results. Population distribution*. Population Census. www.stat.ee

Tench, R., Fawkes, J., & Palihawadana, D. (2002). Freelancing: Issues and trends for public relations practice. *Journal of Communication Management*, *6*(4), 311–322. https://doi.org/10.1108/13632540210807143

The World Bank. (2022). *Mobile cellular subscriptions (per 100 people) - European Union*. Data. Mobile Cellular Subscriptions. European Union. https://www.worldbank.org/en/home

Travers, J., & Milgram, S. (1969). An Experimental Study of the Small World Problem. *Sociometry*, *32*(4), 425. https://doi.org/10.2307/2786545

Türkeş, M. C., & Vuță, D. R. (2022). Telework: Before and after COVID-19. *Encyclopedia*, *2*(3), 1370–1383. https://doi.org/10.3390/encyclopedia2030092

Watts, D. (2004). *Six Degrees: The Science of a Connected Age*. Vintage.

Wellman, B. (2001). Physical Place and Cyberplace: The Rise of Personalized Networking. *International Journal of Urban and Regional Research*, *25*(2), 227–252. https://doi.org/10.1111/1468-2427.00309

Wilson, R. J. (1996). *Introduction to Graph Theory* (4th editio). Longman. https://www.ptonline.com/articles/how-to-get-better-mfi-results

Zhang, L., & Tu, W. (2009). Six Degrees of Separation in Online Society. *Journal.Webscience.Org*, *January 2009*, 1–5. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Six+Degrees+of+Separation+in+Online+Society#4

Non-exclusive licence to reproduce thesis and make thesis public

I, Lika Zhvania,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

"Exploring Social Networks and Spatial patterns of Information Dissemination in Passive Mobile Positioning Data",

supervised by Ph.D. Anto Aasa and Ph.D. Anniki Puura.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Lika Zhvania

20/01/2023